A Social Service Field Guide to Psychological Testing

MORTON L. ARKAVA, Ph.D.

A SOCIAL SERVICE FIELD GUIDE TO PSYCHOLOGICAL TESTING

By

Morton L. Arkava, Ph.D.

Professor and Chairman, Department of Social Work

University of Montana

1974

Published by the Governor's Crime Control Commission

Department of Institutions

State of Montana

(Under the Provisions of Sub-Grant #736147)

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

PREFACE

The production and consumption of educational and psychological tests have increased dramatically since their development over fifty years ago. Greater numbers of tests are being used to both evaluate and guide individuals and to aid in administrative decisions. Workers in the social services today will confront the results of various kinds of psychological and educational tests, since most case histories in use contain test information.

Because many persons currently practicing in the social services lack basic educational preparation in test use, they tend to misuse and underutilize test information. It is imperative that persons engaged in the delivery of human services understand some simple test concepts for use in effective case management.

This guide was developed to help the social service practitioner utilize test results more rationally and consistently. It is not intended to serve as a comprehensive textbook on psychological test administration, interpretation, or utilization, but rather to serve as a basic guide to those persons who have little or no background in the use and

interpretation of psychological and educational tests. For those who have had graduate-level coursework in psychological testing or extensive in-service training, more advanced texts on the subject are advised.

CHAPTER I

PURPOSES OF TESTING

What is a Test?

Cronbach, a noted authority on testing, has defined a test as " . . . a systematic procedure for comparing the behavior of two or more individuals."[1] Others have defined tests as standardized procedures for obtaining a sample of an individual's behavior. Psychologists and others use tests in order to predict what a person might do or to discover what he could do. Similarly, tests may reveal why a person does certain things. While undoubtedly the answers to these questions would be more accurate if the individuals involved could be observed over a long period of time, this is generally not practical in clinical or industrial settings. Therefore, one must rely on the brief samples of behavior provided by tests. Thus, a test involves both a sample of behavior and a procedure for comparing that behavior with the results obtained by others.

The accuracy of the predictions, from test behavior to actual life behavior, depends upon many factors, including the nature and construction of the test--especially with respect to the concepts of validity

and reliability, the conditions under which the tests are given, and the clinical and social sophistication of the examiners.

The concept of validity refers to the degree to which a test instrument actually measures or predicts specific behavior. For example, if intelligence is of specific interest, a desirable test instrument is one which will give an accurate measure of the concept of intelligence as currently defined and accepted. The concept of reliability refers to the consistency of test results over repeated testings: how closely will an individual's test score, or a test score on an alternate version of the same test, approximate the score he obtains on earlier or later testings with the same instrument? From a statistical point of view, reliability is a necessary condition for validity. These concepts will be explored later in greater detail.

The Uses of Tests

The purposes of tests are many, but generally tests are used to provide information for decision making. Cronbach has suggested that " . . . the value of test information should be judged by how much it improves decisions over the best possible decisions made without the test."[2] He points out that if one desires to predict school grades, the information obtained from a scholastic aptitude test will not provide any greater accuracy than previous school grades. Generally speaking, then, tests provide information that is not easily obtained otherwise.

Test information assists in making two different types of decisions: institutional and individual. The purposes of decision making create the major difference between these two categories.

Institutional decisions. Institutions make decisions according to their goals in contrast to the goals and wishes of the individual. Decisions are made from the perspective of the operation and maintenance of that institution. For example, school personnel typically make institutional decisions concerning possible admission of students into college or special training programs. Test results usually affect those decisions. Similarly, the Army uses tests to assess special aptitudes and skills and to place personnel in special assignments or training programs.

Industries, schools, and social service agencies typically make institutional decisions in similar ways. One such agency, a parole board, often needs to decide upon the possible release of a prisoner. To the extent that parole boards attempt to predict the offender's behavior and to select only those who exhibit the least potential for antisocial behavior, they engage in institutional decision making. In a similar way, probation officers and judges attempt to select good risks for probation. A "good risk" is defined as someone whose potential for repeating his offense is thought to be minimal. In such cases, information is not always complete enough to make an intelligent decision. A test's value lies in its potential to provide greater accuracy in decision making over the best possible decisions made without test information.

Individual decisions. Individual decisions pertain to unique and personal conditions. They are those a person makes about some aspect of his or her own life: the determination of a career, whether or not to enter a special training program or to go to college, the selection of a potential mate. In the social services, individual decisions may be made from the perspective of the person involved. Under certain conditions, for example, the social worker will make a decision for the client.

There are several ways test information can be utilized in individual decision making. Vocational and aptitude tests are commonly used to help people make career choices. Numbers of young people often wonder what career best suits them or offers them the best chance of success. Frequently, vocational interest and aptitude batteries allow them focus on areas of interest with high success potential.

Misuses of tests. Many counselors tend to rely too heavily on test information. As a result they may seriously limit the options available to an individual. The writer has observed a number of persons who sought college preparation in social work simply because their high school counselor told them they had a high score in this area on a vocational interest test. One student decided to enter the field because of her high score on a vocational interest battery and social work was the specific field mentioned. Upon close examination, however, her interests and aptitudes did not support her test score, personal commitments, or her common sense. The writer has also talked with students who

decided not to pursue particular programs in higher education because they had low test scores, despite the fact that in at least two such instances both students had achieved very commendable previous records of academic success.

Research has demonstrated that success in academic programs predicts future success better than test scores. Thus, it would seem that a major decision made on the basis of a test score alone is undesirable.

Tests do not make decisions, they merely provide supplementary information for those who do make the decisions. An institutional or individual decision made on the basis of a single test score alone is a gross misuse and a misunderstanding of the purposes and limitations of tests. Tests merely sample behavior at any given time and place, and as such, are subject to various errors. Consequently, in many cases test scores and interpretations are insufficient tools, not to be exclusively relied upon. The reader is urged to utilize <u>all the available information</u> in making any kind of decision.

NOTES

CHAPTER I

1. L. J. Cronbach, "New Light on Test Strategy from Decision Theory," Proceedings of 1954 Invitational Conference on Testing Problems (Princeton, New Jersey: Educational Testing Service, 1955), pp. 31-32.

2. Ibid.

CHAPTER II

CLASSIFICATION OF TESTS

Tests can be classified in a variety of ways--according to structure, purpose, and method of administration. They may be more or less objective or subjective, highly structured or unstructured, designed for administration to groups or individuals. Tests employed in clinical practice include intelligence tests, achievement tests, aptitude tests, interest tests, and personality tests. There are other "special diagnostic tests" frequently used to assess some particular limitation or potential--such as those especially designed to measure the nature and severity of certain types of reading or learning disabilities as well as those disabilities imposed by organic deterioration or damage. Some tests measure "talents" inherent to artistic or musical productivity. Intelligence tests have probably the longest and most comprehensive history in clinical, industrial, and academic settings, a fact due,perhaps,to the belief long inherent in Western society, that achievement and productivity highly correlate with a concept known as "intelligence." Controversies regarding the actual nature of intelligence have raged for thousands of

years so that conflicting opinions and experimental data occupy many volumes.

Since direct social service practitioners are likely to be most concerned with a test's intended purpose, a classification of tests was developed:

1. achievement and aptitude tests, including intelligence tests;
2. personality and interest tests; and
3. special diagnostic tests.

A considerable amount of confusion exists concerning distinctions between intelligence, achievement, and aptitude. The crux of the argument is whether intelligence as a specific concept can be separated from other factors such as previous learning, achievement, and special kinds of aptitudes. Most theorists today would argue that intelligence tests, aptitude tests, and achievement tests all sample and measure various parts of the same thing. For example, Wechsler--who authored several intelligence tests--defined intelligence as "the aggregate or global capacity of the individual to act purposefully and think rationally and to deal effectively with his environment."[1] Others argue that intelligence is a function of the total personality and cannot be separated from other aspects of the personality. However, Wesman advocates perhaps the most comprehensive and one of the most generally accepted definitions of intelligence in the literature: "Intelligence . . . is a summation of learning experiences."[2] This definition recognizes that when measuring

intelligence the result of many learning experiences and diverse performances are actually sampled. Wesman's definition, by implication, does away with artificial distinctions between intelligence, aptitude, and achievement tests. He contends that all of these devices measure what the individual has learned. The difference in labeling merely signifies the different purposes for which the tests will be used. This can be clarified by considering each of the three separate categories--intelligence tests, aptitude tests, and achievement tests.

Intelligence Tests

Intelligence tests comprise a highly specialized field with a vast body of literature and research surrounding their use. A tremendous variety of intelligence tests are available and in use. Each test reflects the specific definition of intelligence and different personality theory commitment of the author. Some tests only include verbal material, others contain much non-verbal material. Some stress problem solving, while others emphasize memory. Certain intelligence tests result in a single total score, for example an I.Q., whereas others yield several scores or subscores.

Varying emphases lead to different test results. One should expect to find differences in the intelligence test scores of the same person who is examined with different tests. In each test, measures of different kinds of abilities are obtained. Under the circumstances it would be surprising if each intelligence gave us nearly the same test result.

For most purposes, intelligence tests are considered measures of general learning or scholastic aptitude, most useful in predicting achievement in school, college, or training programs.

Aptitude Tests

Aptitude tests also attempt to measure an individual's potential for achievement. However, aptitude tests focus on more circumscribed varieties of achievement than do intelligence tests in determining whether an individual has the potential for achievement in a specifically defined area. For example, an individual's artistic or mechanical aptitude may be measured. Although intelligence correlates to a degree with both aptitude and achievement, studies have shown that high intelligence does not necessarily guarantee astuteness or potential in certain areas. Recent data shows that intelligence as generally defined does not highly correlate with creativity, especially in the artistic sense, as has been supposed. It is not uncommon to observe individuals who appear extremely intelligent in the traditional sense of the word but who simply do not seem to possess or have developed certain aptitudes. Witness the college professor or physician who is a whiz in the classroom or operating room but who is helpless when faced with an ailing carburetor.

An aptitude test uses a sample of behavior to predict future performance in some specific occupation or training program.

In general use are two major types of aptitude tests: (1) broad-range aptitude test batteries to sample general aptitudes, and (2) specific

aptitude tests to sample special aptitudes such as music, mathematics, and art.

The most widely used broad-range batteries are the Differential Aptitude Tests (DAT), for high school students, and the General Aptitude Test Battery (GATB) currently utilized by the United States Employment Service. In addition, a myriad of multi-score aptitude test batteries exist.

Often aptitude tests are employed in selecting individuals for jobs, for admission to special training programs, or for scholarships. Primarily, the tests predict an individual's potential for achievement in specific occupations or endeavors.

Achievement Tests

Achievement tests, although in many ways similar to intelligence tests, are generally designed to determine what an individual has actually achieved in a certain area of endeavor. They are used to measure a person's present level of knowledge or competence in, for example, courses like mathematics, science, reading, chemistry, etc. Many achievement tests, unlike other types, are not standardized but are produced locally. For example, teachers normally develop achievement tests to determine mastery of the course material. Thus, an achievement test examines a person's success in past or present study; in contrast, aptitude tests forecast success in some future study. Achievement tests, most widely used in academic settings, are usually

reported in the form of grade levels or similar measures of comparison.

Personality and Interest Tests

Personality and interest tests focus on what a person typically does or might do in a given situation. What personality and interest tests measure, in contrast to intelligence or achievement tests, is far less clearly defined. Here a tremendous number of different terms describe similar kinds of things--terms like adjustment, personality, temperament, interest, preferences, values, attitudes all describe similar, broadly defined attributes. It is difficult to say what a specific personality test score means, even after having given the matter careful consideration.

Personality tests. Clinical psychologists and others interested in the prediction of human behavior have long favored personality tests. They realize that those patterns of behavior usually referred to as "personality" have a strong influence on what we do. Personality, for example, can largely determine how people characteristically use or direct their intelligence and special creative aptitudes. Indeed, personality "deficits" or distortions lead to little constructive use of one's talents. Thus, an assessment of personality is vital to those who attempt to help a person channel his or her efforts toward constructive vocational or social use.

Most personality tests rely on vaguely defined scores and scales used inconsistently from one author to the next, based on an underlying rationale not always specified.

Few people agree on a standard classification of personality tests. However, at least three different types of tests are in general use:

1. The objective test batteries not directly subject to clinical interpretation for initial scoring. The Minnesota Multiphasic Personality Inventory (MMPI) is one example of this particular test.
2. The less commonly used situation test measures performance in complex life-like situations or simulated situations and tests special kinds of abilities involving overall responses of an individual to specific situations. Industry commonly uses it to test leadership abilities. When a person is given a group and a specific task to accomplish, he or she is observed in the process of completing the task.
3. Projective tests are designed to elicit subjects' responses to an ambiguous stimulus such as a picture or an inkblot. The individual's response is interpreted and scored on the assumption that the way he organizes and responds to unstructured or ambiguous stimuli indicates the way he organizes and responds to the world around him. Responses are assumed to be projections of the subject's unconscious

wishes, attitudes, and values. The scoring method is similar to the psychoanalytic method of dream interpretation.

Typical projective tests in wide use are the Rorschach (inkblot) and the Thematic Appreception Test (TAT).

Personality tests are primarily used to predict the future behavior of individuals in both general and specific situations. They commonly aid in predicting post-institutional adjustment for persons released from prisons, hospitals, and schools, or in predicting the likelihood of marital success or job performance. Test reports typically contain terms such as anxiety, ego, libido, cathexis, sublimation, etc. A great deal of controversy surrounds the use of various personality tests, especially regarding their validity and reliability. In general, projective tests are not uniformly accurate in predicting the behavior of individuals in either a specific or general situation, but are accurate, given extreme individuals and extreme situations. Although it may be fascinating, a projective test may prove disappointing if used to accurately predict behavior in a way that might be useful to most practitioners.

Interest tests. Interest tests are specific personality tests used mainly in vocational and educational guidance. They are difficult to separate from aptitude tests, but come under the general category of personality tests because they are directed toward such things as predicting a person's potential satisfaction with a given type of work. The two most widely used interest inventories are the Strong Vocational Interest Blank and the Kuder Preference Record (see example below).

Mr. Williams' scores on the Kuder Preference Record indicate that he is highly interested in science, computational activities, and clerical work. These interests are at the 95th, 91st, and 87th percentiles respectively. He demonstrates moderate interest in art and mechanical areas also. The latter interests are at the 75th and 70th percentiles. Training areas he may wish to consider then are: computer programming, computer technology, x-ray technology, laboratory technology, drafting, mechanical drawing, computer systems analyst, electronics technology, radar technology, chemical standards work, industrial standards work, bookkeeping, accounting, printing, etc. As noted earlier, his intellectual level and academic preparation are quite sufficient for him to be successful in a four-year college or technical program.

Interest tests generally rely on self-reporting techniques and are designed to sample both leisure time and work-related activities given specific personality aspects in the area of personal likes and dislikes. They are used to determine the amount of preference a person displays for one activity over another. For example, the inventories typically sample reading interest by asking people if they would prefer to read about adventure, business, science, or romance. Another example, the California Occupational Preference Survey, samples eight interest categories: science technical, outdoor, business, clerical, linguistic, aesthetic, and service.

Although the interest inventories are considered separately for analysis, they are generally regarded as special personality measures

used specifically to predict occupational, vocational, and educational adjustment. However, for purposes of classification, we may regard them as personality measures that fall under the subcategory of objective testing devices. Almost all of the interest batteries rely on objectively scored testing methods based on standardized methodologies.

Specific Diagnostic Tests

A widely diverse group of tests developed for highly specific purposes tend to defy classification. Most such tests were developed to measure specific abilities or disabilities. Some tests diagnose cerebral pathology such as brain lesions or other organic abnormalities. There is disagreement as to how much these tests actually measure underlying pathology or primary causation as contrasted to possible poor learning conditions. For example, the Bender Visual Motor Gestalt Test, sometimes regarded as a test for the diagnosis of possible brain damage, may also be considered a straight-forward ability test. This test requires the subject to produce various geometric designs using a pencil and paper. The way in which they go about achieving this task is subject to various scoring procedures. Most examiners agree that the Bender Test (BVMGT) is basically a performance test since the examinee is affected by previous learning. However, there is indication that the Bender does some rough screening for identifying persons with possible brain damage.

NOTES

CHAPTER II

1. David Wechsler, The Measurement and Appraisal of Adult Intelligence, 4th ed. (Baltimore: Williams and Wilkins, 1958), p. 7.

2. Alexander G. Wesman, "Intelligence Testing," American Psychologist 23 (1968) : 267.

CHAPTER III

BASIC TEST CONCEPTS

A knowledge of some basic testing concepts, including their construction and utilization, is central to understanding the limitations of various tests. Two concepts constitute the criteria used for judging a test in its totality: reliability and validity.

Reliability

Reliability refers to the consistency of measurement of any test. A test cannot measure anything well unless that something is measured consistently. It is important to realize that although a test measures things consistently it may not measure the desired characteristic.

In Chapter I, tests were defined as samples of behavior. Because they are, they will show variation from sample to sample--that is, we may expect differences in the behavior of an individual from one testing situation to another. Reliability of a test is measured by the extent to which results vary from sample to sample. It is necessary to obtain a high degree of reliability in test results to ensure confidence in that test and to achieve validity.

Factors affecting reliability. A number of variation sources affect the reliability of tests. They are:

1. Test length. Assuming that fatigue does not become a major factor, a longer test is more likely to be reliable than a shorter test.
2. Time between tests. The length of time between two testings will affect reliability. The shorter the time between the two tests the more likely it is that the re-test will be similar.
3. Irregularity of testing conditions. Changes in conditions from one testing situation to another will affect the test reliability. Failure to follow specific directions for giving the test may reveal a considerable amount of difference on the scores obtained from the same test taken by the same individual at different times. Extreme differences in physical conditions--overly heated, uncomfortable test rooms, or poor lighting conditions--will also affect reliability. Other factors such as the examiner's responses, racial differences between the examiner and subject, moods, illness, cheating by the examinee, etc., may threaten test reliability.
4. Scorer error. When tests are not scored objectively, or the details of scoring ignored, unreliability results. Objective tests reduce the possibilities for scorer error. Tests designed to elicit subjective responses require special

training for scoring. Indeed, many score errors occur as a result of the examiner's inexperience.

Determining reliability. Two basic procedures are used to establish test reliability: the test/re-test procedure and the alternate form procedure. The test/re-test procedure involves testing and re-testing the same individuals at different times using the same instrument. Test/re-test results are usually reported as a reliability coefficient which represents the degree of agreement between the two measures.

The alternate form procedure involves using two different tests designed to measure the same thing. Alternate form tests are used when the examiner believes that exposure to one test will contaminate the later responses of an individual if he or she is tested again with the same instrument. In other words, a person may have learned what to expect from a set of specific test questions, therefore, influencing his or her second response to the same set of questions. In this case, an alternate form of test may be developed to test similar attributes. Alternate form reliability is also reported as a reliability coefficient representing the degree of agreement between the two measures.

In addition to reliability determined by the actual construction of the test, it is imperative that the reliability of test scoring be controlled. Scorer reliability may be an important factor for some tests--especially projective tests such as the Rorschach. Scorer reliability can generally be highly developed by providing standardized training

for scorers and by comparing the scoring of several examiners for the same test. It may also be reported as a numerical figure which is usually referred to as an interscorer reliability ratio.

Validity

In order to make a statement about the general validity of a test, it is essential that the user of any psychological test information determine what kinds of decisions he or she is going to make regarding the use of that test. In its broad sense, validity denotes the extent to which a test measures or predicts that for which it was designed. In other words, validity is the most basic and perhaps the most important single attribute of the test for it must do what it is designed to do. If a test is supposed to predict occupational success, the extent to which it does so may be said to be a measure of its validity. However, it is important to recognize that psychological tests may have a high degree of validity for one purpose but almost no validity for other purposes.

Various measures of validity, used in psychological testing, include face validity, content validity, predictive validity, and construct validity. Although a great deal has been written about validity measures, for practical purposes most social service practitioners will primarily concern themselves with predictive validity.

Predictive validity, also called empirical or criterion validity, is established by determining how well a test predicts performance

against a specific criterion. A test's validity is determined by operationally defining what the test should do and what outcomes it can predict: The test's success in predicting that outcome is the extent to which the test proves valid. Thus, if a practitioner uses an instrument to screen people for discharge from a correctional institution, how well that instrument predicts specific aspects of post-institutional adjustment, such as recidivism, determines the test's validity. Again, validity is determined by a specific definition of what the test should do. The same test, for example, might prove ineffective in identifying potential salesmen for an automobile agency.

Schools establish predictive validity by using intelligence tests to predict potential achievement. Scores obtained on specific intelligence tests are compared with grades earned in school. In a similar way the predictions of occupational preference tests are validated by comparing ratings by individuals and employers at a later date.

A number of factors influence predictive validity.

1. The specific criteria used to establish validation may vary from study to study with different scores obtained from each criteria. Therefore, it is necessary to carefully consider which criteria is most important for the decision at hand.
2. Some tests are defined more specifically than others in terms of what they are intended to do. For example, easily identified criteria such as school grades can validate a

scholastic aptitude test but it is very difficult, if not impossible, to establish an acceptable criteria for an anxiety scale or a value scale. Where such difficulty in defining criteria related to the intended results of the test exists, high validity cannot be expected.

Practitioners should keep in mind a general, useful rule of thumb: For a test to have any utility, it must provide accurate information that can help to predict behavior. Thus, the less specific the objectives of the test, the less useful it is in predicting behavior.

Predictive validity is generally reported as a numerical figure called a validity coefficient; a measure of validity achieved by computing a coefficient of correlation between the test and a criterion.

CHAPTER IV

BASIC STATISTICAL CONCEPTS

In addition to understanding reliability and validity as essential concepts underlying test construction, the consumer of test information should know how test results are typically reported.

Norms

The familiar expression, "How are you?" and the response, "In relation to what?", best expresses what norms are all about. That is, they provide those standards against which a given value is compared. Norms may be used to determine how well a person does in comparison to other people.

Many people in the social services view test results simply in the form of a raw score. Without more information, it is impossible to use the test results to make a productive decision. For example, the raw score of 65 may mean that 65 test items were answered correctly. If there is no norm for comparing that score with other people's responses, one can attach no meaning to that score. A set of norms is imperative to understand the meaning of raw test scores.

Cronbach has defined a test as "a systematic procedure for comparing the behavior of two or more persons."[1] In spite of the philosophical difficulties one may have with making comparisons, psychological testing does just this. A norm is nothing more than an average score for a specified group used to make comparisons between individuals and groups. On some tests, for example, the performance of persons in a specific geographic location is compared with the performance of persons nationally.

It must be perfectly clear to the social service practitioner just how an individual's test results compare with specific responses from other groups. Who or what a person is compared to makes a great deal of difference. Consider the following example. Jane Sloe, a seventeen-year-old inmate in the state correctional institution for girls, received a raw score of 163 on a vocabulary test designed to predict academic success in collegiate programs. At this point, her probation officer must decide whether to release her by September 1 so she may enter college. Her raw score of 163 means that she did as well or better than:

99 percent of the residents of the state school for girls;

87 percent of the twelfth-grade students in Capital City;

83 percent of the entering freshmen at State University;

75 percent of the graduating seniors at State University;

96 percent of the custodial and treatment staff at state school for girls;

15 percent of the faculty in the English Department at State University.

Although Jane's absolute performance remains unchanged, the impression of how well she has done may differ considerably as the norm groups change. Admittedly, this illustration is extreme, but it does point up the importance of specifying the norm group to which one compares a person.

Thus, to fully understand a norm group one must gather as much information describing the norm group as possible and determine how the person tested differs from it. In viewing any norm group, consider such important variables as age, sex, previous education, socio-economic background, ethnic membership, and occupation. In other words, one should use the most appropriate norm group for the individual examinee and the situation involved.

Publishers supply norm information on most standardized tests, especially educational achievement tests. Most test publishers routinely report norm information and specify if they will make available both local and national norm information. In addition, standard test references also report norms. Social service practitioners should use this rule of thumb regarding norms: <u>the more information provided describing the norm group, the more accurately you can assess to what extent a given individual resembles the norm group</u>.

Test manuals usually provide broadly based or "national" norms. When using such norms it is important to get more detailed information about the groups used to establish these norms. Most of the norm groups will not comprise an entire population; they sample what the test constructors think of as the relevant population. To establish norms, they divide the relevant population into subgroups that appear in the sample in proportion to their numbers in the population. Ideally, those individuals who comprise the sample from each subgroup are selected randomly. Frequently, test constructors subdivide populations according to such characteristics as rural-urban residence, age, sex, race, socio-economic status, religion, and geographic region. Sometimes they seek to establish from a specific stratified population of people what they consider to be normal performance ranges. In some cases, however, they may leave out certain elements of the population in the original norming groups, thus making the specific test irrelevant for use on that population. For example, one of the most frequent cirticisms of intelligence tests is that adequate samples of American Indians were not included in the original norming group. Such tests, like the Wechsler Intelligence Scales, may make a poor basis for comparing the performance of American Indians to other segments of the population.

Social service practitioners, then, must focus on the detailed description of the norm group's relevant characteristics. Furthermore, when subgroup differences are known to be related to test performances,

it is important to report separate norms for the subgroups. A case in point involves the effects of early child-rearing practices and development in multi-lingual homes. A test that focuses on the development of English vocabulary, standardized on a population of midwestern school children, may not be a valid basis for comparing the performance of southwestern Chicano children who come from Spanish-speaking homes. Again, it may be desirable and even necessary to establish separate norms for people from a similar population before meaningful comparisons are made. One must remember that how accurately a norm group represents the population to be tested is more essential than the absolute size of that norm group. True, the larger the sample the more stable the statistics based on the sample, but a representative norm group of moderate size is more useful than a large, poorly defined group.

Measures of Position

Numbers which tell us where a score value stands within a set of scores are measures of position. There are two commonly used measures of position: rank and percentile rank.

Rank is the simplest description of position. It designates the highest, the next highest, the third highest, and on to the lowest--a simple way of describing the position of a person or a score with respect to a distribution of scores. However, it has a major limitation: its interpretation depends on the size of the group. It is generally used in an informal sense such as designating a person's standing in their high school

graduating class; but the meaning of graduating first in a class of three in the Polaris, Montana High School is not as clear as graduating first in a class of 5,000.

Percentile rank states a person's relative position within a defined group. Thus, a percentile rank of 97 indicates a score as high or higher than those made by 97 percent of the people in that particular group. Percentile ranks are one of the most widely used measures of position for reporting test scores, especially on scholastic achievement tests. Although easily understood and commonly used, percentile ranks have a major limitation; they are based on the number of people with scores higher or lower than the specified score value. Percentile ranks tend to obscure all information about those scores' distribution and the absolute differences in raw scores achieved by individuals. However, this information can be regained by focusing on other measures of variability and central tendency.

Measures of Central Tendency

A measure of central tendency is a representative common denominator for a set of scores. Three common measures of central tendency are in general use: the arithmetic mean, the crude mode, and the median.

Mean. The mean, or arithmetic mean, is nothing more than an average. To arrive at the mean, all the scores are added up and divided by the number of scores.

Median. The median is the midpoint of an array of scores. It is the point above which and below which 50 percent of the scores fall. The median is determined by simply ordering the scores from the lowest to the highest and selecting the middle score in the range of scores represented. For example, if there are five scores present, as follows, 1-3-6-9-12, the middlemost score in this distribution, or the median, is 6. The median is the score's position with respect to others and has very little to do with the absolute value of that score.

Mode. The mode is the most frequent score occurring in a distribution. Thus, for the scores 1-5-5-2-5-3-5, the mode is 5. This is one of the crudest measures of central tendency and is used only for rough estimates.

Measures of Variability

Measures of variability describe the extent of score dispersion in a particular distribution and the degree to which scores vary from each other. It is important to know about variability measures in order to compare the score of a given person or group of persons with the dispersion that is logically or reasonably expected. Common measures of variability include range, semi-interquarterile range, and the standard deviation.

Range. Range arrives at a rough measure of variability by identifying the two most extreme scores, i.e., the highest and lowest

score. Thus, the range is simply the difference between the highest and the lowest score.

The semi-interquartile range. The semi-interquartile range describes the dispersion represented by the middle half of a distribution. In other words, the semi-interquartile range represents the distance between the twenty-fifth and the seventy-fifth percentile on a distribution. It is used in conjunction with the median as a measure of central tendency, especially where the test achieves atypical and highly unexpected distributions.

Standard deviation. Standard deviation is perhaps the most widely used, most dependable measure of variability because it fits mathematically with other statistics and thus becomes the basis for a number of other statistical measures, including standard scores, deviation I.Q.s, T scores, and z scores. The standard deviation is the square root of the mean of the squared deviations from the mean (of a distribution). The standard deviation is generally represented either by the Greek symbol σ or the letter s. Thus $s = \sigma$ = standard deviation.

The standard deviation is used to make interpretations of the variability of scores in a distribution. For example, the standard deviation has known characteristics. In a normal distribution of scores, about 34 percent of those scores lie between the mean and a point that is one standard deviation on either side of the mean. Thus, 68 percent of the scores on that distribution will be dispersed between a point lying one standard deviation below and one standard deviation above the mean. In

other words, we know that in a normal distribution we can expect approximately 68 percent of the population to fall within one standard deviation ± of the mean.

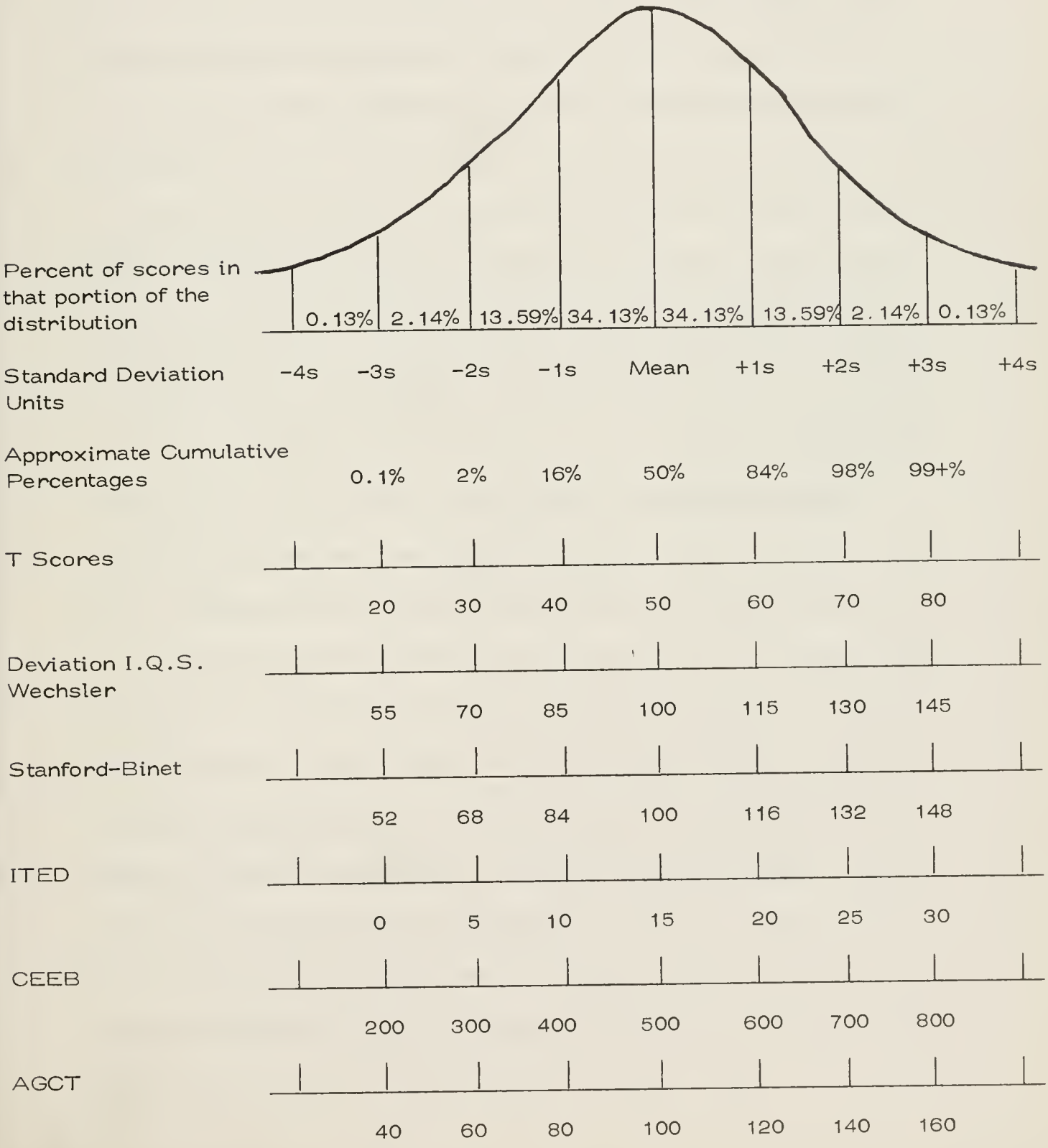
You will find knowledge of this dispersion useful because if you know the standard deviation of a test and a person's score you can use this information to estimate how that person compares with others who have taken the same test. For example, the standard deviation on the Wechsler Adult Intelligence Scale is 15 and the mean is 100. Thus, if we find a person who scores at 145 on the Wechsler, we can translate that information as follows: he has scored three standard deviations above the mean or higher than 99 percent of the people who have taken the examination. This calculation was achieved by adding the number of standard deviations above the mean to the percentage of cases included below the mean (the bottom 50 percent). The standard deviations can be converted into percentages of cases by reference to standard tables such as found in Table I.

Measures of Correlation

Measures of correlation determine how much two or more variables relate to each other. A correlation coefficient is an index number which expresses, in numerical values which range from .00 (no relationship) to +1.00 (perfect relationship) or to -1.00 (negative relationship), the degree of relationship between two or more variables. The higher the number expressed as a coefficient of correlation, the more clearly related the two variables--and one must remember that negative correlations are as important as positive correlations.

TABLE I

COMPARISON OF SOME STANDARD SCORES



Correlation coefficients are utilized in testing to report measures of validity and reliability. Validity studies use them to express the degree of relationship between the test scores and certain criterion values. Reliability studies use them to express the degree of relationship between scores for both test/re-test and alternate forms of reliability.

Two of the several methods used to compute correlation coefficients are the popular Pearson Product Moment Correlation Method and the Spearman Rank Difference Correlation Method. Any elementary statistical textbook contains these and other methods for calculating correlation coefficients and for determining the reliability of those coefficients.

Inferential Statistics

Inferential statistics are sometimes called probability statistics because these measures tell us how much confidence may be placed in descriptive statistics--those numbers used to describe the actual results that people achieve on tests. Generally, inferential statistics are reported as "probability values." They indicate how likely the results obtained on a given test would occur by chance alone. Of the few inferential statistics referred to in test literature, perhaps the most important is the standard error of measurement. The standard error is a statistic used to estimate how likely a specific test score will diverge from the true score achieved by a person. In other words, the standard error indicates how much a person's score would vary if he or she were examined

repeatedly with the same test. This standard error of measurement is one way of expressing a test's reliability.

Other inferential statistics commonly used include the χ^2 (Chi Square) test and the Fisher's t. For further information regarding inferential statistics, a basic textbook on probability statistics should be consulted.

Raw and Standard Scores

The direct numerical report of a person's test performance is called the raw score. This score may represent the number of questions answered, the time required to complete the test, or any other numerical value representing test performance. Raw scores are easily misunderstood because they do not include a basis for comparison. Testers generally convert them into standard scores which are then reported.

Standard scores, derived from raw scores, are used as part of a scoring system which usually offers other information so that all scores provide a basis for comparison. They are used to report test results in almost all intelligence and achievement tests. Remember: (1) most standard scores are based on the properties of the normal curve, and (2) standard scores generally include such information as the mean and the standard deviation of the distribution. They are extremely useful because they permit comparisons between tests that have similar types of scoring systems. The basic standard score is the z score--a method of comparing scores on one test with scores on other tests.

Following is an example of the use of standard scores in common psychological tests. The I.Q., or Intelligence Quotient, used around the turn of the century as a way of measuring the intellectual capacities of people, was computed by determining the ratio between mental age and chronological age. Although this particular method was used in the original Stanford-Binet Individual Intelligence Test, it is now nearly obsolete. Instead, the Stanford-Binet and other tests such as the Wechsler have converted to a standard score known as a deviation I.Q.

The Wechsler Adult Intelligence Scales (WAIS) established the deviation I.Q. by utilizing six verbal subtests and five performance subtests, each of which yield a raw score, establishing different norm groups to cover different age ranges from 16 to 64 years. A raw score distribution resulted for each of the eleven subtests. Each subtest raw score was then converted into a standard score with a mean of 10 and a standard deviation of 3. The sum of all the subtests provides an overall raw score which is converted to a standard score with a standard deviation of 15 and a mean of 100. Tables are provided for the conversion and to adapt for different age groups. Thus, the WAIS user now only needs to consult the appropriate table to find the I.Q. value which corresponds to the sum of the subtest scores.

In a similar way, the 1960 revision of the Stanford-Binet was converted to adopt the deviation I.Q. so that the standard deviation would be consistent from age to age. Previously it was noted that the standard

deviation for different ages in the 1937 revision of the Stanford-Binet differed by as much as 8 I.Q. points. Thus, the Stanford-Binet evolved to have a mean of 100 and a standard deviation of 16. The way in which the raw scores are converted to scaled scores and thus the deviation I.Q. is very similar to the procedure employed in the Wechsler tests.

Another commonly encountered test which is based on a standard score is the College Entrance Examination Board (CEEB), administered by the Educational Testing Service. The CEEB was standardized on a population of college applicants in 1941. Using the scores based on the applicants of 1941, the testers have developed the CEEB as follows. The mean of the test is 500, the standard deviation is 100. With this information in mind, the test user can estimate the relative position of any person's given score. For example, a score of 800 on the CEEB is three standard deviations above the mean. This means that the examinee achieved a higher score than over 99 percent of the population on which the test was standardized.

Other commonly used standard scores include the Stanine Score, the T Scale Score, the C Scale Score, the Sten Score, and the Iowa Test of Educational Development (ITED) Score. Although a number of other types of standard scores are used, Table I will give the reader a general idea of how the commonly used scales compare with each other.

Ratio Scores and Placement Scores

Because the term "I.Q." can mean several different things, it is frequently misused and misunderstood. The original I.Q. concept developed by Terman was a ratio type I.Q. found by the formula I.Q. = $100 \times \frac{MA}{CA}$, where MA is mental age figured from an intelligence test and CA is the examinee's chronological age at the time of testing. The rationale of the ratio type I.Q. is widely understood but this type of score has many limitations for it implies that mental age units are of equal size, which is not verified by research evidence. Ratio I.Q.s work reasonably well if the examinee's age is between five to fifteen. Outside of these approximate limits, they tend toward invalidity. This is because the differences (in mental growth) between the ages of five and six, for example, are much greater than between the ages of fifteen and sixteen or between twenty-five and twenty-six. Unlike chronological age units, mental age units are not equal. Thus, when interpreting an MA or ratio I.Q., the social service worker must be cautious.

Another type of commonly used ratio score is the educational quotient. It resembles the mental age concept and is used to estimate a person's minimal performance in comparison with other people who perform in educational settings. Dividing an educational age score achieved on a test (EA) by the chronological age and multiplying by 100 will result in the educational quotient (E.Q. = $100 \times \frac{EA}{CA}$). Educational quotients have the same advantages and limitations as ratio I.Q.s. However, they are only used with school-level achievement tests.

The most common score used in reporting performance on standardized achievement tests for school children is the grade placement score which resembles the age scores. It is found by determining the average score of school students at a corresponding grade placement. Grade placement scores are usually stated in tenths of a school year. For example, 6.2 refers to the second month of grade six. This approach assumes that children learn relatively uniformly throughout the school year but that no learning occurs during the summer vacation--an assumption not necessarily true. Grade placement scores, however, are generally used to estimate where a person should be placed in school-related work.

NOTES

CHAPTER IV

1. Cronbach, Essentials of Psychological Tests, op. cit., p. 21.

CHAPTER V

LIMITATIONS OF TESTS

Supplementary Measures

The full assessment of a person's abilities, disabilities, and various personal qualities requires a progressive type of approach. Since a test is simply a behavior sample to be regarded cautiously, it follows that test scores are minimal estimates of behavior and abilities. Broad spectrum tests like the Wechsler Intelligence Scales are initial steps to assessing general ability. When a social service worker must make an important decision, he or she should move through progressive stages of assessment, using test information that deals with specific intellectual, perceptual, and/or cognitive factors. Existing tests can only provide clues and rough estimates regarding an individual's abilities and capacities. REGARD THE RESULTS OF ANY ONE TEST CAUTIOUSLY.

Test Construction Limits

One of the major difficulties in deciding how to best use the various kinds of test results springs from the uncertainty over what tests actually measure. Some tests are designed to measure verbal learning and

abstractions, while others assess manual skill potentials. The Wechsler and the Stanford-Binet were originally constructed to measure both performance and verbal factors. Yet many psychologists agree that the weighting of the Stanford-Binet is more toward measuring verbal ability than are the Wechsler tests which strike a more even balance between verbal and performance items. However, a person's achievement on either verbal or performance tests partially reflects previous learning experiences. Thus a number of other factors, all related to previous exposure to similar materials, are important determinants for assessing test limitations.

Factors such as membership in specific geographic groups are related to mastery of subject content. For example, urban residents are exposed to a more diverse range of stimuli than are rural residents. A rural resident may define an elevator as a grain storage facility, while an urban resident may describe an elevator as a device used to move people up and down in a building. If this were a test question, i.e., define an elevator, standardized on an urban population, the rural answer might be judged unacceptable. Such a question will not fairly measure the rural resident's potential ability to acquire knowledge. Similarly, many achievement and ability tests do not fairly sample the potential abilities of members of various ethnic groups.

Effects of Culture

For a test to be truly fair, all of the examinees should have had an equal opportunity to acquire the needed background. There have been many attempts to construct culture-free or culture-fair tests--those that supposedly do not depend on previous experience--but most social scientists believe that experience affects all behavior.

A number of cultural factors affect test performance. Previous training experiences influence outcome. The Zuni Indians are taught cooperation rather than competition, and so the performance of a Zuni Indian on a competitive test may reflect this teaching and could vary considerably from a person reared in a culture which stresses competition. Some tests may exhibit a cultural bias simply because the examinee is not familiar with the testing materials.

Additional factors that affect test scores include the sex and race of the examiner. Carkhuff and Pierce have reported that the race of both the examiner and the tester appear to have a significant effect upon the outcome of clinical interviews.[1] Unless the culture and communication patterns of the group tested are thoroughly understood, the examiner, no matter how unprejudiced or objective, may not obtain maximum results in the test procedure. Sensitivity to all aspects of a subject's behavior is essential for acquiring a fair test result. This type of sensitivity does not develop through academic efforts, but rather through prolonged, intimate contact with the specific ethnic group.

Most tests used for individual mental testing do not truly represent some of the ethnic groups in the United States. For example, since the Stanford-Binet standardized on white, middle-class subjects, it primarily measures verbal ability and generally reflects this culture; subjects from other ethnic groups often do poorly on this scale. The WAIS and the WPPSI include black subjects in the norming group, but not other ethnic groups. Furthermore, since the WISC includes only white students in the standardization sample, the scoring criteria used in establishing correct and incorrect responses discriminate against the responses made by Chicanos and American Indians. Thus, when these standard instruments for determining the ability of ethnic minority members, especially American Indians, rural residents, Chicanos, and others not represented, are used they should be interpreted with extreme caution and regarded as minimal estimates of ability.

When test results are used to evaluate the potential performance of a minority group member, the purpose those results serve should be constantly kept in mind by the social service practitioner. Although a test may discriminate against members of a specific minority group, this will not necessarily diminish the test's predictive validity when used to determine success in, say, a school program. If you want to obtain an estimate of a minority person's chances of succeeding in a training program, a standard intelligence test may prove a valid predicting instrument.

Other Limitations

Keep in mind, always, that the performance of any individual on a test is simply a sample of behavior, and as such, performance in any one test is a minimal estimate. Naturally, the more measures of similar variables the more reliance one can place on such estimates, but any contradictory scores on two or more tests that measure similar attributes should be noted. If such a contradiction exists, seek professional interpretation of the differences--at least three different variables may be involved, including differences between the tests themselves, differences in the individual or the group being tested, and differences in the conditions of test administration from one test to another.

NOTES

CHAPTER V

1. R.R. Carkhuff and R. Pierce, "Differential Affects of Therapist, Race, and Social Class upon Patient Depth of Self-Exploration in the Initial Clinical Interview," Journal of Consulting Psychology 31 (1967): 632-35.

CHAPTER VI

HOW TO MAKE A TEST REFERRAL

Most social service personnel should make test referrals as a normal or routine part of their work. The best test referrals identify the specific information sought.

Since testing should have as its major purpose the provision of useful information for decision making, the psychologist who selects a test that will provide such information needs to know what kinds of decisions the social service worker intends and, in addition, will need to know what kinds of information are already available about the subject. Following is a suggested guide for making test referrals.

Suggested Guide for Test Referrals

1. Reason for referral--what kind of information do you want and what kinds of decisions will you try to make.
2. Description of the subject, including:
   a. age;
   b. sex;
   c. education;
   d. occupation and employment history;

e. ethnic membership and experience; and

f. note any special disabilities, handicaps, or physical abnormalities.

3. The results of any previous testing, if available--including testing dates, scores, and test names.

4. A brief history of the subject's involvement with the agency.

5. A brief statement of any case-management plans you have for the subject.

Given this information, a competent examiner should be able to select the tests that will best provide the information you need. The examiner should also be able to interpret the tests' results in a way that is useful in making specific case management decisions.

Some Hints for Dealing With Psychologists

If your test report is in a form that is difficult to use in decision making, ask the examiner for a consultation or interpretation. But remember, what you get out of such a meeting largely depends on the questions asked. It is advisable to key your questions to specific decisions about the subject. For example, will Johnny get through college at State University? Or, is there a possibility Johnny may commit suicide?

Although no examiner can answer either question with certainty, he/she can provide some information about the probability of either event occurring. Ethical examiners will also provide the necessary explanations about the limitation of the instrument.

Some test reports may be confusing because the examiner used special tests which the test results do not explain. It is always appropriate to ask the examiner to explain the purpose of all tests given.

You should also prepare people referred for psychological tests for the actual examination procedure. Such preparation should include an explanation of why the referral was made and a description of what they should expect in the testing situation--how long it will take, where it will be done, etc. You can obtain this information from the examiner on request.

It is not generally appropriate to request the examiner to administer a specific test. Leave test selection up to the examiner unless you can make a special case to justify an exception.

Many test reports contain much technical jargon. You can always ask examiners to explain all terms you do not understand in a test report. Technical jargon is meant for other psychologists and does not generally convey a great deal of meaning to many test information users.

Psychological examiners have no magical powers. The same information coming out of a psychological examination might well come from others in everyday situations. People who have known the subject over a period of time and have observed him/her in different situations can often tell you more about the subject than most examiners. Remember that tests only sample behavior; what happens in real life also fully indicates what to expect from that person.

CHAPTER VII

SOME COMMONLY USED TESTS

Differential Aptitude Test (DAT)

The DAT battery, originally published in 1947, is currently available in two forms. High schools use it for counseling students in grades eight through twelve. The eight tests measure aptitudes which previous research had found important in guidance.

The DAT primarily provides a standardized procedure for measuring boys' and girls' multiple aptitudes for educational and vocational guidance. It yields separate scores from eight subtests plus a score resulting from a combination of two of the eight subtests. The eight tests are: Verbal Reasoning, Numerical Ability, Abstract Reasoning, Clerical Speed and Accuracy, Mechanical Reasoning, Space Relations, Language Usage (which includes Spelling and Language Usage) which deals with grammar. The tests require six to thirty minutes of working time, plus additional time for directions. Thus, each of three sessions need eighty minutes. Except for the clerical test, tests are not timed.

Both test/re-test and alternate forms of reliability determination have achieved highly acceptable reliability figures.

Validity has primarily been established by attempting to match test performance with later course grades. In this respect, predictive validity has been high enough to demonstrate correlations ranging from .70 to .80. In general, however, the predictive validity for the subscales instrument is not very high, usually around the area of .50. The best overall predictor of grades is the combination score reported for verbal reasoning and numerical ability.

The sampling of over 50,000 students from 195 different schools in 43 states, representing all major geographic areas in the United States, established DAT norms. (There are separate norms for boys and girls and also for Fall and Spring Semester testing.) These norms are expressed as percentile ranks and stanines. Remember: Although the DAT predicts success in coursework and grades reasonably well, it does not adequately predict vocational success. Therefore, use it with considerable caution when advising students on career selection. The DAT should be used with other instruments for best results.

Goodenough-Harris Drawing Test (Draw-A-Man Test)

This test was designed for children five to fifteen years of age to evaluate intelligence by analysis of the child's drawings of a man and a woman. It can be used as an initial screening test, a rapid way of gaining an impression of a child's general ability levels and as a means of estimating the mental ability of children for whom the usual verbal tests of ability are inappropriate.

The test booklets provide three spaces for the child to produce drawings: one for the drawing of a man, one for a woman, and one for a self-portrait. The examiner asks the child to draw the very best picture possible of a man, a woman, and himself or herself. The child is cautioned to make a whole person, not simply a head and shoulders view. Although the test is not time limited, the child usually completes it in ten to fifteen minutes. Tests may be administered either to groups or to individual children.

The test contains fairly explicit scoring directions. It has, according to research, relatively high coefficients of interscorer reliability (approximately .90), but tests for test/re-test reliability only range from a test/re-test reliability of .94 for a one-day interval between testing to .65 for a three-year interval between testings. Most test/re-test reliability coefficients for other tests range between .60 and .70.

The Draw-A-Man Test's validity has been primarily demonstrated by correlations with scores on other tests. Correlations with the earlier forms of the Stanford-Binet range from .30 to .74, and range similarly in other tests: reports show correlations of this test and the WISC between .40 and .50.

Samples of 300 children at each age level from 5 to 15 years, selected as representative of the population of the United States according to father's occupation and geographic region, established norms for these scoring scales. (The test manual reports standard score norms which have a mean of 100 and a standard deviation of 15.)

Limitations. Because, on this test, the standardized sample of population held only a few critical variables constant (such as father's occupation and geographic region) the norms may not be useful for special ethnic groups such as American Indians or residents of extremely rural areas. Also, this test's validity has been established primarily by comparison with scores on other tests.

Thus, the Draw-A-Man Test may only generally indicate the likelihood of a child scoring well on another test, and may be invalid as a predictor of potential performance in training programs. In summary, the social service practitioner should regard this test with caution for it only roughly estimates intellectual ability. Other measures of general intellectual ability should supplement it.

Other Drawing Tests

Although almost every art media, technique, and type of subject matter has been investigated in the search for significant diagnostic clues, special attention centers on drawings of the human figure. The Machover Draw-A-Person Test is a well-known example. In this test, the examiner provides the subject with a letter-size sheet of paper and a medium-soft pencil and tells him/her to simply "draw a person," or--to young children--"draw somebody" or "draw a boy or girl." Upon completion of the first drawing, the examiner asks the subject to draw a person of the opposite sex from the first figure. While the subject draws, the examiner notes comments, the sequence in which parts are drawn, and

other procedural details. An inquiry may follow this drawing in which the subject is asked to make up a story about each person drawn "as if he were a character in a play or novel." During the inquiry a series of questions elicits specific information about age, schooling, occupation, family, and other facts associated with the characters portrayed.

Qualitative judgements, involving the preparation of a composite personality description from the analysis of the many features of the drawing, and considering the absolute and relative male and female figures' size, their position on the page, quality of lines, sequence of parts drawn, stance, front or profile view, position of arms, depiction of clothing, and background and grounding effects--all of these make up the scoring of this test. Omission of bodily parts, disproportions, shading, amount and distribution of details, erasures, symmetry and other stylistic features produce special interpretations. Each major body part, such as head, individual facial features, hair, neck, shoulders, breast, trunk, hips, and extremities, is regarded as significant.

The interpretive guide to the Draw-A-Person Test contains sweeping generalizations, such as "disproportionately large heads will often be given by individuals suffering from organic brain disease" or "the sex given the proportionately larger head is the sex that is accorded more intellectual and social authority." But no evidence supports these statements. The guide also refers to a "file of thousands of drawings" examined in clinical context and a few selected cases are cited for

illustrative purposes, but no systematic presentation of data accompanies the published test reports.

Validation studies of this test by other investigators have yielded conflicting results. Attempts to develop semi-objective scoring procedures which utilize rating scales or checklists have met with little success. The test may succeed more with children and other relatively naive subjects than with sophisticated adult groups. Although it appears to differentiate between seriously disturbed persons and normals, its discriminative value within relatively normal groups is questionable. Research on the Draw-A-Person Test has been inadequate largely because of failure to cross-validate.

The House-Tree-Person Projective Technique (H-T-P) devised by Buck, has aroused considerable interest as witnessed by the number of relevant research publications. In this test, the subject is told to draw as good a picture of a house as possible, then the same for a "tree" and a "person." Meanwhile, the examiner takes copious notes on time, sequence of parts drawn, spontaneous comments by the subject, and expressions of emotion. Oral inquiry, including a long set of standardized questions, follows completion of the drawings. The examiner analyzes the drawings both quantitively and qualitatively, chiefly on the basis of their formal or stylistic characteristics.

In discussing the rationale underlying the choice of objects to draw, Buck maintains that "house" should arouse association concerning the

subject's home and those lived with; "tree" should evoke associations pertaining to life goals and ability to derive satisfaction from the environment in general; and "person" should call up associations dealing with interpersonal relationships. Some clinicians may find helpful leads in such drawings when considered jointly with other information about the individual case. The elaborate, lengthy administrative and scoring procedures described by Buck appear unwarranted in light of the highly inadequate nature of the supporting data.

Minnesota Multiphasic Personality Inventory (MMPI)

The design of the MMPI provides an objective assessment of some of the major personality characteristics that affect personal and social adjustment. The scales provide a measurement for the personality status of literate adolescents and adults together with a basis for evaluating the acceptability and dependability of each test record. Nine scales were originally developed for the test's clinical use and were named for the abnormal conditions on which their construction was based. Since they have proved meaningful within the normal range of behavior, these scales are now referred to by their abbreviations--Hs (hypochondrasis), D (depression), Hy (hysteria), Pd (psychopathic deviate), Mf (masculinity-feminity), Pa (paranoia), Pt (psychoasthenia), Sc (schizophrenia), and Ma (hypomania)--to avoid possible misleading connotations. Development of these test items has produced a number of other scales: Si (social

introversion) is commonly scored, as well as three validating scales: L (lie), F (validity), and K (correction). A wide variety of untrained personnel can administer this inventory, however, a thoroughly trained clinical or educational psychologist should interpret the results.

One can expect test subjects sixteen years of age or older with at least six years of successful schooling to complete the MMPI without difficulty. When an individual is specifically referred for testing, one can generally ascertain beforehand whether the MMPI is appropriate for use and thus avoid the embarrassment that would arise from failure during the actual administration. The full-scale edition of the MMPI requires the subject to give a true or false response to 566 separate questions (see Table I). The raw scores thus obtained are converted to a kind of standardized score called a T score on which the MMPI profile and code are based. The test items are presented either in a card form for individual use or in a booklet with a separate answer sheet for individual examination or large-scale group testing programs. Such a profile provides a scale for clinical comparison of the relative "strength" of various personality trends.

Clinicians who use the MMPI usually tend to emphasize one particular scale of the nine. The MMPI should not be evaluated on the basis of one scale alone but rather on the pattern of scores for the entire nine scales including the validity indicators. The test affords an infinitely large number of patterns. Thus, although scorers may often

feel that they have seen some given pattern a number of times before, almost no exact duplicates exist.

Although originally thought of as an aid to psychiatric diagnosis and evaluation, the MMPI has been used in many different settings and validated against hundreds of different criteria. The rapid rise of these tests' non-psychiatric application has stimulated a substantial growth in new scales and scoring procedures.

Reliability and validity research on the MMPI are not entirely convincing. Validity studies do not show high correlations between MMPI profiles scores and actual psychiatric diagnoses, although the instrument was initially developed for this purpose.

Indeed, the available categories of psychiatric diagnoses are subject to criticism since it is questionable whether or not the MMPI actually achieves its intended objectives when used strictly clinically. But where the MMPI is used to screen large populations, such as military recruits, college students, or business executives, it serves as a reasonably reliable, general screening device. It is most useful in identifying those persons who achieve extreme scores on the subscales--thus identifying those who require further study.

The use of the MMPI requires professionally trained, experienced, and sophisticated practitioners, because of the complexity of the personality characteristics of the inventory, the meanings of the scales, and the way in which the scales relate to each other in predicting behavior.

The original MMPI was standardized on a sample of about 700 normal visitors at the University of Minnesota Hospital (ranging in age from 16 to 55 and representing a cross section of the Minnesota population) in contrast to some 800 clinical cases (from the Neuro-Psychiatric Division of the University of Minnesota Hospital).

The test/re-test method determines the reliability for the MMPI. Reliability results show that the coefficients of correlation vary considerably with different subscales. The test/re-test reliability coefficients range from .46 to .93 with the majority lying between .70 and .90--a fairly high degree of reliability.

The predictive method, which compares the scores obtained on special scales with clinical diagnoses for newly admitted psychiatric patients determined validity for the MMPI. In approximately 60 percent of the cases this method predicted the corresponding clinical diagnosis.

Sample test reports. The client's responses to the MMPI indicate a dependent, immature, impulsive, demanding woman who attempts to exploit and control others. She seems able to maintain relationships only with those she can keep in subservient positions. Probably her fear of abandonment creates this fear. Unfortunately, she seems not to recognize the alienating effect of her domineering tactics. Her imperious manner and her repeated demands will drive any away from her except

those even more emotionally disturbed than she. She is a very angry woman who seems especially resentful toward men. While she pretends to heterosexuality, she may spend a great deal of her time trying to prove this through sexual acting out, repeated love affairs, etc., probably because she has an amorphous sexual identity. While not psychotic, apparently she is poorly controlled, disorganization-prone, moody, and hypertensive. Her obesity is probably a function of anxiety. She eats to ward off the loneliness and to control the gnawing emptiness of feared abandonment. She has a personality disorder, perhaps a passive-aggressive personality of the aggressive type. She needs individual psychotherapy and will probably not lose weight nor be able to stabilize vocationally without this. She will benefit best from a reality-oriented, problem-solving approach although she might make use of "insight." She will probably have a stormy relationship with any therapist.

Minnesota Counseling Inventory

An effort to adapt the previously discussed Minnesota Multiphasic Personality Inventory for use with normal high school students and college freshmen led to the development of the Minnesota Counseling Inventory. Many of the 355 true/false items of the latter inventory came from the MMPI, and several other scales have a close resemblance to the MMPI scales. With norms based on over 20,000 high school students tested in ten states, this test provides scores in seven areas designated as: Family

Relationships, Social Relationships, Emotional Stability, Conformity, Adjustment to Reality, Mood, and Leadership. The "Conformity" scale has a strong resemblance to the MMPI Pd scale and "Adjustment to Reality" similarly resembles the Sc scale. Also, two vertification scores exhibit similar traits to the MMPI validity scales. The comparison of random samples of students with groups nominated by teachers as outstanding examples of the quality tests, validated the total scores on the different scales. Test reliability established by split-half and re-test procedures is at an acceptable level. But the seven area scores are not as distinct as their titles imply. Only counselors familiar enough with its construction to evaluate its complex scores should use this inventory.

Otis Self-Administrating Test of Mental Ability

An early test that has been widely used in personnel screening on a group basis is the Otis Self-Administrating Test of Mental Ability. This test also helped to develop the basis for the highest level norms for Otis Quick-Scoring Mental Ability Tests used as an academic screening device from the early grades through high school level. Industry uses the Otis Self-Administrating Test of Mental Ability for screening applicants for such varied jobs as clerks, calculating machine operators, assembly line workers, and foremen and other supervisory personnel. A number of validation studies have checked the Otis against an industrial criterian, most of which have demonstrated that the scores of the applicants compare with actual job performance creating significant validity coefficients. In

semi-skilled jobs, the Otis Test correlates moderately well with success in learning the job and ease of initial adaptation. It does not, however, correlate highly with subsequent job achievement. This would be expected for routine jobs, but also holds true for high-level, professional jobs since it discriminates poorly at these upper levels.

General Aptitude Test Battery (GATB)

The U.S. Employment Service produced this battery. Throughout the country it helps to guide people seeking work. State employment services give these tests as well as other non-profit agencies whose personnel have been trained in the use of the test by the Employment Service. High school juniors and seniors often take them through a cooperative plan which makes the results available to both the high school counselor and the employment service. Versions of the tests have been prepared for a number of foreign countries.

The Employment Service constructed the test to help guide persons into suitable work. Each of the thousands of jobs in the modern industrial world has its own aptitude requirements. When an employer asks for referrals of potential employees, he wants applicants likely to succeed. The U.S. Employment Service working with state agencies, therefore, conducts studies of the psychological characteristics of particular jobs and accumulates information on the meanings of a test score. The following illustrates the small sample of the occupations studied: assembler of dry cell batteries, aircraft electrician, teacher, x-ray

technician, nurses' aide, sheet metal worker, baker, cook, spot welder, comptometor operator, corn husking machine operator, knitting-machine fixer, food packer.

Predictions for such jobs takes us far beyond the academic ability and reasoning ability which predominate most aptitude tests. The diversity of occupations rules out the possibility of devising a separate aptitude test for each job. For guidance, a limited number of diversified tests are needed which everyone can take and which can be linked together in various combinations to predict success in different situations. With this end in view, the current GATB measures nine distinctive factors:

G - General reasoning ability (a composite of tests entitled Vocabulary, Three-Dimensional Space, and Arithmetic Reasoning)

V - Verbal aptitude (Vocabulary)

N - Numerical aptitude (Computation, Arithmetic Reasoning)

S - Spatial aptitude (Three-Dimensional Space)

P - Form perception (Tool Matching, Form Matching)

Q - Clerical perception (Name Comparison)

K - Motor coordination (Mark Making)

F - Finger dexterity (Assemble, Disassemble)

M - Manual dexterity (Place, Turn)

No other similar test exceeds the efficiency of the GATB. Each of its paper-pencil tests takes about six minutes. The psychomotor tests require even less working time but several minutes are used for demonstration practice. The entire battery can be given in two and one-quarter hours. The simple procedures allow trustworthy administration of the tests by relatively untrained testers to subjects who have limited education or poor command of English. The psychomotor tests are designed so that each subject leaves all the materials as they were found--ready for the next subject.

This marked speeding of nearly all the GATB subtests may reduce their validity for many purposes, especially if the person has some reading deficit, is upset by tests, or has taken few tests. But since the U.S. Employment Service had access to workers in all areas of the country, all types of industry and agriculture, and most occupational levels, it could obtain a highly representative normal sample. It drew 4,000 cases from the records on hand to form a group which properly represented all occupational, sex, and age groups in proportion to census data.

Test results are reported as standard scores with a mean of 100 and a standard deviation of 20. Extensive research has demonstrated good reliability and validity. Validity does vary between and among different occupations. The social service practitioner should use the GATB in conjunction with the U.S. Employment Service's Dictionary of Occupational Titles.

An example of a GATB test result is as follows: Mr. Smith's scores on the Intelligence, Verbal Aptitude, and Numerical Aptitude sections of the General Aptitude Test Battery indicate that his achievement is far above average in each of these categories. His Intelligence score is at the 99th percentile as compared to the general working population, while his Verbal and Numerical scores compared to the same population are both at the 96th percentile. All of his scores on the remaining aptitudes of the battery are at the 75th percentile or above including his scores on the Manual and Finger Dexterity Form Board. It is apparent, then, that Mr. Smith is intellectually capable of undertaking technical or college training in any of the occupational aptitude patterns covered by the GATB. That is, he has the intellect and dexterity necessary to handle any of the many occupational categories listed from Occupational Aptitude Pattern 1 through Occupational Aptitude Pattern 35 inclusive. His interest profile from the other tests in the battery suggests that he may wish to consider any of the following occupations listed in Occupational Aptitude Pattern 1: physician, civil engineer, highway engineer, etc. Under Occupational Aptitude Pattern 2, he may wish to consider training as a pharmacist, cost accountant, tax accountant, or statistician. Appropriate occupations from Occupational Aptitude Pattern 3 are teacher, survey worker, group worker, or caseworker.

Strong Vocational Interest Blank (SVIB)

One of the most widely used interest tests is the Strong Vocational Interest Blank, first published in 1927. The Strong contains questions on hundreds of activities both vocational and avocational. Most of the 400 items require a "like-indifferent-dislike" response to activities or topics: biology, fishing, being an aviator, planning a sales campaign, etc. Because research has demonstrated that the majority of men in a particular occupation have roughly similar interests, the Strong assumes that a person having a typical occupational group pattern will find satisfaction in that field.

The Strong determined the interest pattern for a profession by giving the questionnaire to successful members of that particular profession and by comparing the responses of the group with those of men of similar age selected randomly from the whole range of occupations ordinarily entered by college men. A weighted scoring key assesses how closely the subject's interests correspond to those of the professional group. On each item, the percentage of men-in-general for each answer was compared to the percentage of men-in-the-occupation giving the answer. Engineers dislike "actor" more commonly than other men; therefore, response D, or dislike, is assigned a positive weight in the engineers' scale. "Liking to be the author of a technical book" (a significant indicator of engineering interests) is especially common among engineers, thus acquiring a weight of plus three. In contrast, 40 percent of artists respond "like" to "actor." So

the artist's scale weights "actor" at plus two for like, zero for indifferent, and minus 1 for dislike.

Occupational scores convert into letter grades ranging from A to C. Seventy percent of successful men in the occupation fall into the A group on that scale. Interests of a person who falls below B plus are quite different from those of the bulk of the occupational group. Only 2 percent of the men in the occupation fall as low as C.

The test has available a great number of keys for male occupations and a woman's blank which can be scored for a number of occupations typically entered by women. The Strong contains items varied enough to predict almost anything, and a new key can be made for any vocational or specialized interest group. Strong keys can score not only vocational interests, but also it can score answers which men give more frequently than women, for example, and create a "masculinity-feminity key."

Extensive research has demonstrated considerable predictive validity for this instrument. Strong demonstrated that interest scores obtained by college undergraduates predicted their occupations of eighteen years later. His interest scales successfully differentiate members of an occupation from the population in general and the occupations from each other. Given the amount of research on both the reliability and validity of the Strong, it is reasonably assumed to be one of the best occupational predictive instruments available.

However, caution seems indicated in interpreting the results for both the Strong and the Kuder for a number of studies have demonstrated that examinees can fake these inventories: Examinees, told to attempt responding in a way that they thought life insurance salesmen would, generally succeeded in making themselves appear like life insurance salesmen. In other words, if a person suspects what characteristics are being screened, this person can fake a response. However, evidence does not suggest that people in general fake their responses but that most people are genuinely interested in their test outcomes. Below is an example of a Strong test.

The results of the Strong Vocational Interest Blank indicate a client highly interested in religious activities, social service, and music, as well as public speaking, business management, art, teaching, mathematics, technical supervision. His general interests show a similarity to those men successful as music teachers and music performers, but also similar to those of credit managers, chamber of commerce executives, business education teachers, social workers, YMCA staff members, rehabilitation counselors, public administrators, physical therapists, and librarians. Surprisingly, in view of his stated vocational aspirations, his interests do not parallel those of computer programmers. With this in mind, he must revise his planning. While computer science is not an altogether inappropriate career choice, the client would probably be happier in a career more oriented toward administration and dealing with the public. Computer science may provide an opportunity for this,

especially if supplemented by general business and/or management courses. He may also wish to consider a business curriculum, college, or trade school. Counseling also appeals to him so various types of social work may be feasible. But since he shows ardent interests in music, he should explore the possibility of becoming a music teacher or performer.

Stanford-Binet Scale

The Stanford-Binet scales for measuring intelligence (since 1937 known as the Stanford-Binet Scale) has gone through several revisions, all of them using a common principle: the average capacities of children of various ages represent differences in degrees of brightness along with differences in levels of development. Thus, knowledge of intellectual performance levels of typical children of a given age facilitates comparison with any specific child by comparing his/her score with the average. The principal criterion employed by Binet and Simon in the standardization and age-placement of tests was: any item successfully completed by two-thirds to three-fourths of a representative age group of children of a given age was designated as "average" performance for that age group, and their ideal standard placed the test at a year-level passed by 75 percent in that age group.

The following procedure describes the method of scoring the Stanford-Binet Scale. The examiner selects a starting point in a range of tasks where the subject can pass all items. This is called the "basal year." The examiner then proceeds upward in the scale until the subject fails all

items, a level called the "terminal year." Each item carries specified credit in terms of months contributing to the mental age score. These credits, added to the age value of the basal year, total the mental age. For example, assume a basal year of six; then, three test items passed at the seven-year level give additional credit of six months, two passed at the eight-year level give further credit of four months, but all failed at the nine-year level. Thus, the subject's mental age is six years, ten months.

The 1937 revision of the 1916 scale differs in many details from its predecessor (unsatisfactory items were eliminated and new ones added), but it shares the essential and basic conception. It has two equivalent forms, L and N, each of which contain 129 test items as compared with the 90 items in the first Stanford-Binet. The 1937 scale extends downward to the level of age two and upward through three levels of "superior adult" (known as superior adult I, II, and III) thus increasing its usefulness.

From the ages of two through five, this scale provides groups of test items at half-year intervals and thus obtains more accurate and highly differentiating test results. The half-yearly intervals are possible because the mental growth rate proceeds most rapidly in the earlier years creating more rapid periodic increments susceptible to testing.

Although the 1937 scale like that of 1916 relies predominantly on its verbal character, it does provide performance and other non-verbal materials, especially through the age of four years. Performance materials demand

the subject to do something--build a pattern or make a design with blocks or fill in a form built with variously shaped blocks. Other non-verbal materials include such activities as copying a geometric figure, completing the picture of a man, discriminating between forms, etc. In all these, the child must use verbal ability inasmuch as he/she must understand verbal directions. In these tests, verbal ability can also be a factor if the child knows the names of the objects or geometric figures and this knowledge helps the manipulation or classification of them.

Since the 1937 scale was standardized on only American-born, white, primarily urban subjects, it is also extremely verbal and thus additionally culturally loaded. Though this test is still used with children, the Wechsler Intelligence Scales have largely replaced it.

Vineland Social Maturity Scale

This scale, designed for use with individuals from infancy to the age of thirty years, models itself on the construction and standardization of the Stanford-Binet scale.

Unlike many other scales, this one is based upon a well-defined rationale and systematic construction. It groups behavior items at age levels as in the Stanford-Binet; these items represent progressive maturation and adjustment to the environment in the following categories:

Self-help - reaches for nearby objects (age 0-1)

Self-direction - buys own clothing (age 15-18)

Locomotion - walks about room unattended (age 1-2)

Occupation - helps at little household tasks (age 3-4)

systematizes own work (age 25 plus)

Communication - makes telephone calls (age 10-11)

Socialization - demands personal attention (age 0-1)

advances general welfare (age 25 plus)

Examiners score items after interviewing someone well-acquainted with the subject or the subject himself. Then, a social age is obtained which is divided by chronological age, yielding a social quotient (S.Q.).

Although this social maturity scale highly correlates with intelligence test results (about .80), the author maintains that its content and rated function are distinct enough for use in the study of an individual's general behavioral development, since social age provides a procedural basis to guide the care and training of an individual.

While the scale aids in diagnosing the normal population as well as the mentally deficient, it was first conceived to facilitate the diagnosis of mental retardation. Primarily it differentiates between mentally retarded individuals who can conduct their personal and social life with greater independence and the mentally retarded who are socially inadequate.

Clinics widely use the Vineland Scale with children and adolescents. And, in addition, it is a valuable device for interviewing and counseling both parents and children.

Thematic Apperception Test (TAT)

Commonly referred to as the TAT, this projective personality test consists of thirty picture cards plus one blank card. An examiner uses the cards in various combinations depending upon sex and age; some are used with all subjects and others with only one sex or age group. The examiner uses only twenty total pictures with any subject which are usually administered in two test sessions, ten pictures at a time.

Examinees are told that the TAT tests imagination. They are to make up stories to suit themselves and are assured no right or wrong responses exist. The examiner shows pictures one at a time, giving simple instructions and asking the subject:

1. to tell what he/she thinks led up to the depicted scene, how it came about;
2. to give an account of what is happening and the feelings of the characters in the picture; and
3. to tell what the outcome will be.

The test has no time limits and an examiner encourages the subject to continue for as long as five minutes on a picture. Sometimes the examiner uses an interview to learn the origins of the stories, especially associations to places, names of persons, dates, specific and unusual information are sought. This is an important aspect of the process because it enables the examiner to clarify stories' meanings. For instance, a boy ten years of age made up a surprisingly large number of stories

dealing with death. The interview revealed these as normal responses: his father was an undertaker and they lived above the funeral parlor.

Although the TAT uses pictures more structured than an inkblot, they possess enough ambiguity to allow wide latitude for individual differences in responses. The TAT is, like the Rorschach, a projective method. Murray designed the TAT to elicit "drives, sentiments, and conflicts" by analysis of the story produced by the subject. He based the test upon the principal that when interpreting an ambiguous social situation, one is apt to reveal aspects of one's own personality that would not or could not be admitted otherwise because they are unconscious. The subject, while absorbed in the picture and attempting to construct an appropriate account of it, is off guard and becomes less aware or quite unaware of himself/herself in the situation. In creating stories based upon somewhat vague pictures, the subject utilizes and organizes content of unique personal experiences. The examiner regards everything the subject says as having meaning. From these stories, the skilled examiner/interpretor draws inferences regarding the subject's personality traits and their organization. The limitations of other projective devices also limit the TAT. A number of different elaborate and special-purpose schemes allow scoring of the TAT, but they show little uniformity in procedure for analysis of test results and few clinicians report the specific system in use. Thus, comparisons between examiners are often impossible.

Unless one of two objective, specific scoring systems is used along with a specially trained scorer, reliability for the TAT is generally low.

Validity research has not demonstrated the TAT's practical use. It helps little in predicting behavior and thus is of little value in decision making. However, it has been useful for research in achievement motivation. Below is an example of the TAT.

The client's responses to the TAT indicate a chronically anxious, impulsive person who becomes flighty, disorganized, and hypermanic under stress. He avoids close relationships because he can relate only in a superficial, exploitive way. He wishes those stronger than himself would take care of him and thus he may go to rather great lengths to make people he sees as superior notice him. He has a negative, poorly defined identity. He feels alone, helpless, and unable to function without high anxiety unless involved in a constant frenzy of activity. His feminine interests equip him little to compete with more aggressive peers. While he is not necessarily an overt homosexual, he may be primarily homoerotic in his sexual responses. He fears exploitation and attack. He is afraid of failure and so may not see tasks through to their conclusion. He has many personality deficits and functions in a way which will interfere with constructive achievement in a vocational training program. In fact, his enrollment in a training program should probably be made contingent upon regular psychological treatment. He will respond best to supportive, problem-solving approach and behavior modification techniques emphasizing reward for constructive efforts.

Symonds Picture Story Test (SPST)

The Symonds Picture Story Test is a projective technique designed for the study of the personality of adolescent boys and girls. The SPST is identical to the Thematic Apperception Test except that it uses a different set of pictures especially designed for the study of adolescent fantasy. But, the SPST similarly uses twenty pictures divided into two sets. If both sets are used, the examiner should use one set at a first setting and the second set at a second setting at least a day later (usage has demonstrated Set B the more effective of the two). The examiner individually administers the test in an interview situation which requires about an hour to run through the ten pictures. The author recommends interpreting the results of the test within the context of the subject's life history material secured by casework with psychoanalytic study.

Many of the limitations inherent in the use of any projective device affect this test, and the comments about the use of the Thematic Apperception Test apply fully to the SPST--with the additional observation that the SPST has not been subject to as much research as the TAT. Normative data based on forty cases are available in the manual.

Wechsler Intelligence Scale for Children (WISC)

Examiners frequently use the WISC--an individually administered general intelligence test--to predict academic success or discover intellectual or academic deficiencies which may be interfering with school achievement. Like the Wechsler Adult Intelligence Scale, the WISC

obtains I.Q.s by comparing each subject's test performance with the scores earned by individuals in his or her age group. I.Q.s obtained by successive WISC re-tests always compare the subjects to their own age group at each time of testing. Each person tested is assigned an I.Q. which represents the intelligence rating relative to his or her age. The WISC uses a mean of 100 and a standard deviation of 15, and places I.Q.s from 90 to 110 in the average range. In terms of percentile limits, the highest 1 percent would have I.Q.s of 135 and above, and the lowest 1 percent I.Q.s to 65 and below. The middle 50 percent of children in each age will have I.Q.s ranging from 90 to 110.

The WISC consists of 12 subtests which, like the adult scales, divide into two subgroups identified as verbal and performance. The verbal subtests are: Information, Comprehension, Arithmetic, Similarities, Vocabulary, Digit Span; the performance subtests are: Picture Completion, Picture Arrangement, Block Design, Object Assembly, Coding, Mazes.

In the standardization of the WISC, every subject took all twelve tests, but to shorten the time required for examination the scale has been reduced to ten tests. (The Digit Span in the verbal, and Mazes in the performance part were omitted primarily on the basis of their relatively low correlation with the other tests on the scale, and, in the case of Mazes, the time factor.) One can use all subtests but in this case all twelve tests must be prorated before computing the I.Q.s. Usually a trained clinical or school psychologist administers the test.

An original norming group of only white urban children sampling 1,100 girls and 1,100 boys in 11 age groups standardized the WISC. Its present limitations as a diagnostic instrument for intelligence probably lie in these inadequate sampling procedures--how can the WISC validly test those who radically depart from those in the original sampling? The omission of rural American Indian children in the original sampling procedures sharply limits this test's usefulness with them. Shifts in population distribution, general levels of education, and increased use of information dispersement by mass media means may further invalidate WISC results on present day populations.

Though I.Q. tests can and will be administered to so-called disadvantaged groups, the social service practitioner must remember to interpret the results with great caution, as most items on these tests are culturally biased in one way or another. Remember: These tests minimally estimate "intellectual ability" and the results should be supplemented by intelligence tests when the meaning of the I.Q. test is vague. Use special care in making future predictions on the basis of I.Q. tests alone--especially with young children. Research suggests that the I.Q. does change under certain conditions. Below is an example of a WISC test.

Jimmy Jones is functioning in the bright normal range of intelligence. On the WISC, he achieved a Verbal I.Q. of 112, a Performance I.Q. of 115, and a Full Scale I.Q. of 114. Although Jimmy's relatively high scores on General Comprehension, Similarities, and Picture Arrangement indicate

that he has considerable abstracting ability and that his intellectual potential is quite high, the discrepantly low scores on Arithmetic, General Information, and Vocabulary indicate that he has not been able to make the most of his intellectual potential. Judging from his history and present living circumstances, the discrepancy is probably due to the effects of severe cultural deprivation. The relatively low score on Digit Span also suggests a significant level of anxiety which may be indigenous to test situations. This often appears in children from culturally deprived environments and, of course, adds to the type of school under-achievement that may be reflected in his low scores on the Verbal subtests correlated with such achievement. Certainly, his high scores on Picture Completion, Block Design, and Object Assembly indicate extremely good perceptual-motor functioning. This lends strength to the impression of intellectual functioning that is substantially higher when measured by his overall performance on the Verbal subtests. Probably, his potential lies well within the superior range of intelligence (I.Q. = 120-130).

Wide Range Achievement Test

The Wide Range Achievement Test first standardized in 1936 and revised most recently in 1965, consists of three subtests, each divided into two levels--Level I designed for children between the ages of five years and eleven years and eleven months, and Level II designed for persons from twelve years to adulthood. The three subtests at both levels are:

1. Reading - recognizing and naming letters and pronouncing words.
2. Spelling - copying marks resembling letters, writing the name, and writing single words to dictation.
3. Arithmetic - counting, reading number symbols, solving oral problems, and performing written computations.

Untrained school personnel can administer this test to large groups. The Wide Range Achievement Test has proved valuable in a number of areas:

1. the accurate diagnosis of reading, spelling, and arithmetic disabilities in persons of all ages;
2. the determination of instructional levels in school children;
3. the assignment of children to instructional groups progressing at similar rates and their transfer to faster or slower rates in keeping with individual learning rates;
4. the establishment of degrees of literacy and arithmetic proficiency of mentally retarded persons;
5. the checking of school achievement of adults referred for vocational rehabilitation and job placement;
6. the selection of personnel at various occupational levels for promotion in business, industry, and the National Services; and
7. the selection of students for professionalized technical schools.

Test scores are reported as grade norms and standard scores. Originally the actual mean grade levels of the children in each age group tested established the grade norms. Such an arbitrary score as grade

rating may vary with promotion practices and socioeconomic levels. For example, in 1936, the average person in the norm group obtained a 9.1 grade rating at age 17, but in 1963, the average person obtained a grade rating of 10.8 at the age of 17. This may mean that more people stay in school longer, or that more persons obtain higher grade ratings but not necessarily higher achievement than they did 25 years earlier. Despite these variations, the grade ratings tend to be a rather stable score. The comparability of the old and new grade ratings are striking through nearly all educational levels except the upper ratings. The grade ratings above age 14 are more arbitrary than those below 14.

The standard score in the Wide Range Achievement Test compares to the I.Q. of standard tests. Persons of different ages may receive identical grade scores. For example, a 5.5 grade stands for superior achievement if obtained by a 7-year-old, but represents defective achievement if obtained by a 25-year-old person. The standard score shows whether the grade rating lies above average or below average for any particular age level. The standard scores used are based on the distribution of the grade ratings for each group.

While the Wide Range Achievement Test provides a useful measure of actual achievement, the social service practitioner should not use it alone for it contains multiple and varying reasons for underachievement. For this reason, as with any other test, a lone achievement test score may mislead especially if interpreted by those unfamiliar with the complexities of achievement testing.

Bender-Gestalt

The Bender-Gestalt was designed to test visual motor performance skills. It is used as an aid in assessing perceptual-motor coordination disorders which are often related to organic brain dysfunction. The test has proven somewhat useful in diagnosing various types of retardation, and personality patterns or trends.

The Bender-Gestalt consists of nine different geometric figures, printed on cards, which the subject is asked to reproduce. This basic procedure, called the "copy phase," is in some adaptations of the Bender-Gestalt. For example, the examiner may ask subjects to recall the figures they drew, elaborate upon or change the figures they reproduced.

While an extremely experienced clinician may profitably use this test, it does not lend itself well for "cook book" interpretation or for use by unexperienced examiners, nor should it be used in making final determinations regarding organic brain dysfunction, perceptual-motor deficits, or personality malfunction. A visual-motor task, in testing personality reaction, provides a sample of behavior involving complex functions. Like other so-called "projective" procedures, such complex behavior examples are best interpreted from a consistent theoretical frame of reference. The Bender-Gestalt results only hint at possible brain disorders or personality malfunctions. The theory underlying personality testing through a visual-motor task is that such testing has some special characteristics and possible advantages. The theory notes that probably

styles of perceiving and reproducing figures which are relatively neutral, i.e., have few associations with one's past, tap some personality facets which conscious attempts cannot disguise. A few highly skilled clinicians piece together some good hunches about a subject from the drawings, but since they rely on hunches the Bender-Gestalt remains an experimental instrument, and validation studies prove disappointing despite extensive use. Most social service practitioners would not find Bender-Gestalt test results useful in decision making. After all, how does the social worker use a a test report indicating "suspicion of organicity"? Thus, regard this tests' results with caution. Below is an example of a Bender-Gestalt test result.

This is a record of a 15-year-old girl who is quite inhibited and generally fearful in her behavior. There is evidence of some mild, diffuse organic damage probably occuring in early childhood, perhaps between 4 and 7 years of age, and possibly due to an encephalitic condition. Although she has partially compensated for the intracranial damage, the organic factor still exerts something of a handicap to her adjustment. However, at present, her central problem appears to be neurotic inhibition accompanied by some depression and apathy. The organic factor certainly contributes to her present adjustment difficulties but is insufficient to explain them. At present, her prime means of defense are withdrawal, denial, and isolation. She is quite fearful of rejection in general and is especially fearful of rejection at the hands of those she perceives as authority or parental figures. While she usually tries to conform on a conscious level, she shows fairly

pronounced passive, oppositional tendencies. Under stress, there may be some regression to narcissism and orally dependent behavior. Despite this, she shows some progression and has established some behavioral configurations characteristic of both anal and oedipal periods of adjustment. She desperately needs closeness with people but is fearful of interpersonal relationships and has not developed skills for encouraging these. Rather, she remains in superficial, rather distant relationships, while embellishing these with fantasy. Presently she is moderately depressed, partly because she does not get the attention she needs and partly because of guilt over impulses which she ordinarily inhibits. At this point, she seems to be especially fearful of heterosexual relationships and may suffer from unresolved sexual feelings for her father. Although her major identification is female, her self-image is that of an inadequate person. Characteristically, she remains withdrawn, aloof, and rather "retarded" in her behavior. An estimate of her intellectual abilities measured by her present Bender performance would yield an I.Q. of approximately 75. This rough estimate probably characterizes her present school performance but is not a good reflection of her potential. Despite her attempt to cooperate on the test, there is considerable evidence of marked impairment of intellectual functioning on the basis of neurotic problems. If her neurotic difficulties were resolved, and her severe inhibitions removed, she should probably function within or near the average range of intelligence.

Rorschach

The Rorschach test, or "inkblot," originally developed in 1921 by Hermann Rorschach, has been considerably researched to expand and improve upon its diagnostic virtues and uses. It is used primarily as a personality test based on the "projective" method. The test consists of ten inkblots presented one at a time to the subject. The first seven blots are essentially black and white although blots two and three have smaller red blotches. The last three blots are multi-colored. Typically, the test is individually administered in three phases. During the first, the subject gives spontaneous responses to the inkblots. During the second, or "inquiry" phase, the examiner asks questions to determine how "the characteristics" of the inkblots triggered the subject's response, such as if color, shape, or shading helped the person see what he or she saw. In a later phase (sometimes used), called "testing the limits," the examiner attempts to get additional scoring material, especially if the subject has given extremely unusual responses or has not "seen" the concepts commonly "seen."

The Rorschach scoring system allots five scores to a response. The scoring is determined according to the "area chosen," the content chosen, the form level of the response (this refers to how accurately or arbitrarily, or how definitely or vaguely the response form is conceived), and the "popularity" of the response (whether or not the response is found often or considered extremely rare).

In addition to the Rorschach's complex scoring, a qualitative approach can also add further data. The way a subject approaches the card, the pauses, the difficulties, the apparently extraneous comments can all add further data when interpreted by a skilled clinician.

The theory implies that persons' reactions to these abstract blots will give clues to their reactions to life--one person organizes the blot in minute detail while another may give it a slap dash once over. Is the subject interested in unusual details or in the more common ones? Are colors perceived and reported in the blot description? And so forth.

Although the Rorschach has been used for over fifty years and has been extensively researched to establish its predictive validity, the results have proven somewhat disappointing and uneven. In research experiments, clinicians asked to make a diagnosis based on Rorschach responses alone, without any other data available, could not accurately predict behavior nor diagnose psychiatric disorders. The research indicates that the Rorschach is fickle--sometimes it works while other times it does not. In general, most clinicians agree that the Rorschach has some predictive validity which does better than chance alone. But experienced examiners do equally well by asking subjects direct questions. However, apparently no empirical evidence demonstrates that the Rorschach or any other projective instrument will reliably predict behavior in the day-to-day world. Thus, as a tool in making decisions on practical problems, the instrument is limited.

A number of split-half and test/re-test reliability studies are available on Rorschach protocals. Reported values differ considerably from study to study and for different types of subscores. However, in general, the use of a specific scoring system produces uniformly positive and fairly high correlations in many cases.

The Rorschach, as other projective tests, is a clinical instrument that should give reliable, valid results only when used by persons having both the special technical training and an advanced sophistication in the understanding and application of a specific personality theory. The tests are generally time consuming, both to give and to score, and they are sometimes hard to justify by the results obtained. It is the author's impression that a highly experienced clinician, willing to engage the subject in in-depth interviews, can obtain similar results and perhaps make more meaningful inferences regarding behavior prediction than the Rorschach. Below is a sample test report on the Rorschach.

The Rorschach results indicate that the client is a relatively well adjusted man whose personality conflicts are not sharp enough to cause him chronic or severe anxiety. Indeed, the anxiety he experienced in the present testing situation is probably germane to those circumstances in which he feels he is being evaluated or judged. He fears being found in need--not an unrealistic fear in our society. He is overly concerned about his physical functioning at present, probably due to his history of physical disorder. A sensitive young man responsive to the needs of

others, he is motivated to improve himself and has the impulse control necessary for emotional and intellectual growth. He seems to get along well with others and will probably work hard to do a good job in anything he undertakes. He likes to make friends and seems something of an extrovert. He will probably be popular with his peers, co-workers, etc.

Wechsler Pre-School and Primary Scale of Intelligence (WPPSI)

Published in 1967, this scale is designed to test the intelligence of children from ages four to six and one-half years. The scale includes eleven subtests, ten of which determine the I.Q. score. Eight of the subtests are downward extensions and adaptations of the Wechsler Intelligence Scale for Children (WISC). The other newly constructed three replace WISC subtests that, for a variety of reasons, proved unsuitable. As in the WISC and the WAIS, the subtests group into Verbal and Performance scales from which Verbal, Performance, and Full Scale I.Q.s are found. However, in order to enhance variety and to help maintain the young child's interest and cooperation, the administration of Verbal and Performance subtests are alternated in the WPPSI. Total testing time ranges from 50 to 75 minutes in one or two testing sessions. Abbreviated scales or short forms of the scale are not recommended. The subtests include Information, Vocabulary, Arithmetic, Similarities, Comprehension, Sentences, Animal House, Picture Completion, Mazes, Geometric Design, and Block Design. "Sentences" is a memory test substituted for the WISC Digit Span. The child repeats each sentence

immediately after the examiner orally presents it. This test can be alternatively used for one of the other Verbal tests; or it can be administered as an additional test to seek further information about the child and so it is not included in the total score for calculating the I.Q. "Animal House" is basically similar to the WAIS Digit Symbol and the WISC Coding test. A key at the top of the board has pictures of a dog, chicken, fish, and cat, each with a differently colored cylinder (its "house") under it. The child should insert the correctly colored cylinder in the hole beneath each animal on the board. Time, errors, and omissions determine the score. "Geometric Design" requires the copying of ten simple designs with a colored pencil.

The WPPSI was standardized on a national sample of 1,200 boys and girls in each of six and one-half year age groups from four to six and one-half, where each child was tested within six weeks of the required birthday or mid-year date. The sample was stratified against 1960 census data with reference to geographical region, urban-world residence, proportion of whites and non-whites, and father's occupational level. Raw scores on each subtest are converted to normalized standard scores with a mean of ten and a standard deviation of three within each one-fourth year group. The sum of the scaled scores on the Verbal, Performance, and Full Scale are then converted to deviation I.Q.s with a mean of 100 and a standard deviation of 15. Although Wechsler advises against using mental age scores because of their possible misinterpretation, the

manual provides a table for converting subtest raw scores to "test ages" in one-fourth year units.

Reliability coefficients for the Full Scale I.Q. are acceptably high and consistent with the other Wechsler scales. The manual also provides tables for evaluating the significance of score differences. This data suggests that a difference of fifteen points or more between the Verbal and Performance I.Q. is significant enough to be investigated. Stability over time was also checked in a group of fifty kindergarten children re-tested after an average interval of eleven weeks. Under these conditions, the reliability coefficients for the Full Scale I.Q., the Verbal I.Q., and the Performance I.Q. were satisfactorily high.

The manual reports comparisons with the Stanford-Binet and the WISC. Along with the WISC, the Stanford-Binet correlates higher with the Verbal I.Q. (.76) than with the Performance I.Q. (.66). This group, which was somewhat below average in ability, had approximately the same mean on Stanford-Binet and the WPPSI (91.3 versus 89.6). Similar comparisons at different ability levels are needed.

Owing to its recent publication, little can be concluded at this time about WPPSI's validity and practical use. The procedures followed in standardizing the scale and estimating reliability and validity are of uniformly high technical quality. Both the size and composition of the norm and sample are considerably advanced over the pre-school tests previously available. But observe caution when using any test score

involving young children, for many variables, difficult to control, affect the uncertain procedure of assessing young children.

Peabody Picture Vocabulary Test

The Peabody Picture Vocabulary Test, designed to provide a verbal intelligence estimate through measuring hearing vocabulary, is effective with average subjects, and has special value with certain other groups. Reading is not required, so the scale is especially fair for non-readers, and since responses are non-oral, the test can be used for the speech impaired (expressly the aphasic and the stutterer). It is also used with certain autistic, withdrawn, and psychotic persons. Since neither pointing nor oral responses are required the test can be used with orthopedically handicapped and cerebral palsied persons, and also with some visually handicapped and perceptually impaired persons. Thus, the scale allows for any English-speaking resident of the United States between two years, six months and eighteen years who can hear words, see the drawings, and indicate "yes" or "no" in a manner which communicates. The Peabody Picture Vocabulary Test has a number of advantages:

1. The test has high interest value and thus establishes good rapport.
2. It needs no extensive, specialized preparation for its administration.
3. It is quickly given in ten to fifteen minutes.

4. Scoring is completely objective and quickly accomplished in one to two minutes.
5. It is completely untimed and thus is an ability rather than a speed test.
6. No oral response is required.
7. Alternate forms of the test are provided to facilitate repeated measures.
8. The test covers a wide age range.

The administration of the Peabody Picture Vocabulary Test requires no special preparation other than complete familiarity with the test materials which include giving the test prior to its use as a standardized measure. The examiner must know the correct pronunciation of each of the test words as given in Webster's New Collegiate Dictionary. If all the instructions are completely observed, psychologists, teachers, speech therapists, physicians, counselors, and social workers should be able to give the scale accurately.

Only ten to fifteen minutes are usually required for this untimed test. The scale is administered only over the critical range of items for a particular subject. The starting point, basal, and ceiling vary from testee to testee. The examiner presents a series of pictures to each subject. There are four pictures to a page and each is numbered. The examiner says a word describing one of these four pictures and asks the subject to point to or tell the number of the picture which the word

describes. Subjects are encouraged to "guess" if they do not know which picture best conforms to the meaning of the word presented. The examiner starts subjects at different "picture levels" according to the age ranges specified in the manual, and proceeds forward from the starting point until the subject makes the first error. If the subject does not make eight consecutive correct responses prior to this first error, the examiner returns immediately to the starting point and works backwards (through the next lowest age range) until a total of eight consecutive correct answers are made by the subject. Responses above the starting point--as well as below--are counted in order to establish the basal of eight consecutive correct answers. The examiner then continues testing forward from the point of the first error until the subject makes six errors in any eight consecutive presentations, counting the last item presented as the subject's ceiling. The total score is the number of correct responses. All items below the basal point are assumed correct; all items above the ceiling item are assumed incorrect. To get the total raw score, the examiner subtracts the errors from the number of the last item presented, or ceiling item. By using special tables, the raw score can be converted to three types of derived scores:

1. an age equivalent (mental age);
2. standard score equivalent (intelligence quotient); and
3. a percentile equivalent.

The age norms for converting raw scores on the Peabody Picture Vocabulary Test to mental age scores are given in the manual. Age equivalents supposedly provide an index of the level of a given subject's development. For example, 75 is the mean raw score on Form A for children who have a chronological age of 10.0. Therefore, regardless of subjects' chronological ages, if they obtain a raw score of 75 on a Peabody Picture Vocabulary Test, they supposedly possess a mental age of ten years since their ability to score on this test is like the average 10-year-old's. Approximate grade equivalents derive from age equivalents by the rule of five. Thus, a child with a mental age of 11.0 would have a grade equivalent of six (subtract five from the mental age) indicating an accumulative capacity to achieve at the beginning grade six level. Age norms have a number of advantages. They provide an easily understood index of the subject's developmental level. They are useful in comparing mental age with chronological age, achievement age, social age, and so on. In addition to the age norms, they provide standard score norms which may provide an "index of brightness" for a given child in comparison with other children of the same age. The Peabody was standardized with a mean of 100 and a standard deviation of 15.

Wechsler Adult Intelligence Scale (WAIS)

The WAIS, the adult form of the Wechsler Intelligence Test, is used to assess general and specific intellectual ability for persons sixteen

years and above. The WAIS consists of eleven subtests grouped into a verbal scale and a performance scale.

Verbal Scale

1. Information: Twenty-nine questions covering a wide variety of information that adults presumably should acquire in our culture. An effort was made to avoid specialized or academic knowledge.
2. Comprehension: Fourteen items in each of which the subject explains what should be done under certain circumstances, why certain practices are followed, the meaning of proverbs, etc. These are designed to measure practical judgment and common sense. This test resembles the Stanford-Binet comprehension items but its specific content was chosen to be more consonant with the interests and activities of adults.
3. Arithmetic: Fourteen problems similar to those encountered in elementary school arithmetic. Each problem, orally presented, is to be solved without the use of paper and pencil.
4. Similarities: Fifteen items requiring the subject to say how two things are alike.
5. Digit Span: Orally presented lists of three to nine digits to be orally reproduced. In the second part, the subject must reproduce backwards lists of two to eight digits.

6. Vocabulary: Forty words of increasing difficulty presented both orally and visually. The subject is asked what each word means.

Performance Scale

7. Digit Symbol: This is a version of a familiar code-substitution test which dates back to the early Woodworth-Wells Association Test and has often been included in non-language intelligence scales. The key contains nine symbols paired with nine digits. The subject's score is the number of symbols correctly written within one and a half minutes.
8. Picture Completion: Twenty-one cards, each containing a picture with some part missing. The subject must tell what is missing from each picture.
9. Block Design: This test is reproduced in designs increasing in complexity requiring from four to nine cubes. The cubes or blocks have only red, white, and red-and-white sides.
10. Picture Arrangement: Each item consists of a set of cards containing pictures to be rearranged in proper sequence so as to tell a story.
11. Object Assembly: This test includes a number of pieces to be assembled very much in the manner of a jigsaw puzzle. The subtest includes four pictures to be reproduced including mannequin, hand, profile of a face, and side view of an elephant.

Both speed and accuracy of performance are taken into account in scoring Arithmetic, Digit Symbol, Block Design, Picture Arrangement, and Object Assembly.

The WAIS standardization sample was carefully chosen to ensure its representativeness. The principal normative sample consisted of 1,700 cases including an equal number of men and women distributed over 7 age levels between 16 and 64 years. Subjects were selected to match as closely as possible the proportions of the 1950 U.S. Census with regard to geographic residence, urban-rural residence, race, white versus non-white, occupational level, and education. At each age level, one man and woman from an institution for mental defectives was included. Supplementary norms for older persons were established by testing an "old-age sample" of 475 persons aged 60 years and over in a typical mid-western city.

Raw scores on each WAIS subtest are converted into standard scores with a mean of 10 and a standard deviation of 3. These scaled scores were derived from a reference group of 500 cases which included all persons between the ages of 20 and 34 in the standardization sample. All subtest scores are thus expressed in comparable units. Verbal, Performance, and Full Scale scores are found by adding the scaled scores on the six verbal subtests, the five performance subtests, and all eleven subtests respectively. The manual provides tables which convert these three scores to deviation I.Q.s with a mean of 100 and a standard deviation of 15. However, such

I.Q.s are found according to the specific age group. Thus, they show an individual's standing in comparison with persons of his or her own age level. Deriving I.Q.s separately for each age level compares the individuals with the declining norm beyond the peak age. The age decrement is greater in performance than verbal scores and also varies from one subtest to another. Thus, Digit Symbol, with its heavy dependence on speed and visual perception, shows the maximum age decline. However, on the other performance subtests speed may be an unimportant factor in the observed decline. In a special study on this point, subjects in the old-age sample were given those tests under both timed and untimed conditions. Not only were the score differences under the two conditions slight but the decrements from the 60-64 to the 70-74 age group were virtually the same under timed and untimed conditions.

The WAIS has demonstrated consistently high reliability coefficients through the split-half reliability technique. Validity has primarily been established through demonstrating correlations between test scores and scholastic achievement. The WAIS has also been compared to other instruments for similarity in scores achieved by the same subjects. In all respects the WAIS has demonstrated relatively high correlations. In summary, the WAIS is perhaps the best general adult intelligence test currently available. Following is a sample test report on the WAIS.

The client is functioning within the normal range of intelligence. On the WAIS, she achieved a Verbal I.Q. of 92, a Performance I.Q. of

95, and a Full Scale I.Q. of 92. The client's vocabulary is smaller than average, she thinks in a slightly "scattered" way and has trouble completing tasks that require concentration or the systemmatic organization of intellectual material. There is great variability in her intellectual performance, in fact, and this is typical of the intellectual functioning of those who experience severe anxiety. In this client's case, the results are blocking, inattention to detail, mild confusion, and diminished ability to maintain cognitive set. She works better at structured and unambiguous problems than she does at those requiring her to be organized or to work out novel solutions. The degree of variability in her performance suggests that she would be functioning near the bright normal range had she had better learning opportunities and were she not handicapped by chronic anxiety and emotional difficulties.

Tests for Special Purposes

A variety of tests have been developed for a number of specialized purposes. The following are examples of special purpose tests with references to further information for the consumer.

The Culture Fair Intelligence Test. This is a paper and pencil test developed by Cattell and Cattell, published by the Institute for Personality and Ability Testing. The test is available for three different age levels, ranging from children to adults. The test's purpose is to provide a measure of ability directed at separating the evaluation of natural intelligence from that contaminated or obscured by education. The

Culture Fair Intelligence Test used both the classical I.Q. with a mean of 100 and a standard deviation of 24 and a standard score I.Q. with a mean of 100 and a standard deviation of 16. The best research available on the test indicates that when used in industrial countries similar to the United States the results have been consistent from country to country. In very dissimilar countries, however, the results are significantly different from those obtained with the standardization sample. Extreme caution is urged in interpreting the results of this test for people who come from markedly different cultures. The Institute for Personality and Ability Testing, Champagne, Illinois, offers a manual providing more information about this test.

Tests for the orthopedic handicapped. The Pictorial Test of Intelligence, available through Houghton-Mifflin Company Publishers, requires neither manipulative nor speaking responses. It was designed to assess the general intellectual ability of children between the ages of three and eight and can also be used to test those children who are orthopedically handicapped and cannot respond orally or in writing. The manual provides information with regard to deviation I.Q. norms and mental age norms and percentile norms. Thus, scores may be reported in all three forms. Other tests which have been used with orthopedically handicapped include the Progressive Matrices Test, the Peabody Picture Vocabulary Test, and the Columbia Mental Maturity Scale.

Tests for the hearing handicapped. Several tests have been used to assess the mental ability of people who are hearing handicapped: the Nebraska Test of Learning Aptitude, the Pintner-Peterson Performance Scale, the Arthur Point Scale, and the Point Scale of Performance.

Tests for the deaf include the Point Scale of Performance Tests available in two forms from C.H. Stoelting Co. and from the Psychological Corporation. Both are designed to test persons from five years of age to adulthood. The purpose of the scale is to provide a measurement of the intellectual ability of deaf children, children suffering from reading handicaps, and non-English-speaking children. The test was standardized on about 1,100 public school children from middle-class American homes. Scores are reported in the form of mental age norms and a ratio I.Q.

Tests for the blind. Several standard tests have been adapted for use with blind populations, including the Stanford-Binet and Wechsler scales. The Interim Hayes-Binet Scale is composed of items in forms L and M of the Stanford-Binet which do not require vision. Currently, a special adaptation of the Wechsler scale is widely used for testing the blind. The major adaptation of the Wechsler omitts the performance subtests. The Haptic Intelligence Scale for Adult Blind is also available which was designed to test blind adults aged sixteen and above. This test's results are reported in the form of deviation I.Q.s with a mean of 100 and a standard deviation of 15. The test manual published by Psychology Research in Chicago authored by Shurrager and Shurrager contains further information. The test is also described in Buros' Mental Measurements Yearbook.

CHAPTER VIII

HOW TO LEARN ABOUT SPECIFIC TESTS

Although detailed information is provided here about a number of different tests, this is not intended to be a comprehensive reference guide describing all of the large number of tests available. Following are some standard references which social service practitioners might use to obtain more information about specific tests.

Mental Measurements Yearbook, Oscar K. Buros, editor, Island Park, New Jersey: Gryphon Press, 7th edition, 2 volumes, 1972. Also by the same author Tests in Print, Island Park, New Jersey: Gryphon Press. Currently there are seven editions of the Mental Measurements Yearbook, the latest published in 1972. The Mental Measurements Yearbook lists most of the published standardized tests in print as of the year the book was printed. Those tests not reviewed in the earlier editions are described and criticized by various authorities. The Tests in Print book is a comprehensive test bibliography and index and provides the following information about available tests: the name of the test, the levels for which it is used, the publication date, specialized comments

about the test by various authorities, the number and types of scores provided, the authors, the publisher, and the reference to test reviews in Mental Measurement Yearbook.

Other good books on testing include:

Cronbach, Lee J. Essentials of Psychological Testing. New York: Harper and Row, 1970.

Thorndike, Robert L. and Elizabeth Hogen. Measurement and Evaluation in Psychology and Education, 3rd edition. New York: John Wiley and Sons, 1969.

Robb, George, L. C. Bernardoni, and R. W. Johnson. Assessment of Individual Mental Ability. San Francisco: Intext Ed. Publishers, 1972.

Berdie, Ralph, et. al. Testing in Guidance and Counseling. New York: McGraw Hill Book Co., 1963.

Other sources of information are the test reviews and research in professional periodicals. Journals such as Educational and Psychological Measurement, The Journal of Educational Measurement, The Journal of Counseling Psychology, and the Personnel and Guidance Journal typically carry reviews of some of the more recent published or revised tests.